Grundkurs Statistik für Historiker: T. II, Induktive Statistik und Regressionsanalyse

Thome, Helmut

Veröffentlichungsversion / Published Version
Themenheft / topical issue

Empfohlene Zitierung / Suggested Citation:
Thome, H. (1990). Grundkurs Statistik für Historiker: T. II, Induktive Statistik und Regressionsanalyse. *Historical Social Research, Supplement*, 3, 1-277. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-286012

HSR Supplement/Beiheft No. 3 (1990)

HELMUT THOME

GRUNDKURS STATISTIK FÜR HISTORIKER

TEIL II:

INDUKTIVE STATISTIK UND REGRESSIONSANALYSE

Inhaltsverzeichnis

Inhaltsverzeichnis Teil 1

Editorial Staff: Rainer Metz (Assistant Editor), Thomas Rahlf (Assistant).

HSR was composed by the program *Satz* of *TUSTEP*. (*Tü*bingen *S*ystem of *Te*xtprocessing *P*rograms). HSR was printed by HUNDT-DRUCK (Cologne).

Vorwort zu Teil II

Der zweite Teil des statistischen Grundkurses ist wie Teil I aus Vorlesungsmanuskripten für das Herbstseminar des Zentrums für Historische Sozialforschung entstanden. Er setzt die Kenntnis elementarer Verfahren deskriptiver Statistik voraus, wie sie in Teil I (oder anderen Lehrbüchern) erläutert werden (siehe das hier erneut abgedruckte Inhaltsverzeichnis zu Teil I).

Für Teil II sind die Themen der induktiven Statistik (»Inferenzstatistik«) und der Regressionsanalyse aus zwei Gründen ausgewählt worden. Zum einen beinhalten sie Verfahren, die unmittelbar in der Forschungspraxis des Historikers oder Sozialwissenschaftlers anwendbar sind. Zum anderen werden unter ihrem Dach Konzepte entwickelt, die grundlegend für fast alle »fortgeschrittenen« statistischen Verfahren der Datenanalyse und der Modellkonstruktion sind - wie z. B. Faktoren-und Diskriminanzanalyse, log-lineare Modelle, Strukturgleichungsmodelle, Analyse latenter Klassen, Zeitreihenanalyse, »event-history-analysis«.

Der potentielle Leser oder Kursteilnehmer, der dieses Skript einmal rasch durchblättert, sollte sich nicht durch die vielen Formeln und Gleichungen abschrecken lassen. Es kommt nichts Mathematisches vor, was nicht in der Sekundarstufe II des Gymnasiums behandelt wird; das meiste geht nicht über das Niveau der Sekundarstufe I hinaus. Außerdem habe ich mich bemüht, das, was in mathematischer Symbolsprache dargestellt wird, ausführlich in »Normalsprache« zu erläutern. (Mathematisch orientierte Leser werden vielleicht sogar finden, daß ich an einigen Stellen zuviel diskursiv-sprachlichen Aufwand betrieben habe.) Es ist aber notwendig, sich auch der mathematischen Sprache zu bedienen, um ein Minimum an Genauigkeit zu erreichen und den Leser auf weiterführende Literatur vorzubereiten.

Die Stoffauswahl bezieht sich auf eine Lehrveranstaltung mit 14 doppelstündigen Vorlesungen (plus Übungen). Kapitel oder Abschnitte, die mit einem Sternchensymbol (*) markiert sind, können in einem ersten Lektüredurchgang ausgelassen werden. Sie dienen vorwiegend dazu, die übliche Kluft zwischen Einführungstexten und der einschlägigen Fachliteratur zu verringern.

Die Fußnoten sind kapitelweise durchnumeriert; die Abbildungen befinden sich am Ende eines jeden Kapitels.

Neben den Mitarbeitern des Zentrums für historische Sozialforschung bedanke ich mich vor allem bei Frau Dr. Ilse Costas, Universität Göttingen, und Herrn Dipl.-Soz. Steffen Kühnel, Zentralarchiv für empirische

Sozialforschung, die kompetent und geduldig verschiedene Fassungen des Manuskripts durchgesehen und wertvolle Korrektur- und Verbesserungsvorschläge gemacht haben. Mein Dank gilt auch Herrn Thomas Rahlf, der den größten Teil der Textverarbeitung für die Endfassung besorgt hat.

Köln, Juni 1990 Helmut Thome

Vorwort zu Teil I

Seit Mitte der siebziger Jahre hat die Historische Sozialforschung, die sich an den allgemeinen Methodenstandards einer empirisch-analytischen Wissenschaft orientiert, vor allem unter jungen Historikern, Doktoranden und Studenten zunehmend Interesse und Anerkennung gefunden. Das hat die geschichtswissenschaftlichen Ausbildungsgänge an den Hochschulinstituten der Bundesrepublik wenig beeinflußt. Lehrveranstaltungen über formale und quantifizierende Analysemethoden sind noch immer Ausnahmen.

Das Zentrum für Historische Sozialforschung (ZHSF) führt deshalb seit Jahren »Herbstseminare« durch, in denen neben anderen Inhalten zwei Grundkurse zur statistischen Methodenlehre angeboten werden. Das hier vorgelegte Skript zur deskriptiven Statistik entstand im Rahmen des Lehrprogramms zum »Grundkurs I«. Ein darauf aufbauendes Skript zum »Grundkurs II: Inferenzstatistik und Regressionsanalyse« wird voraussichtlich bis Ende 1989 druckreif vorliegen (siehe die hier im Anhang abgedruckte Themenübersicht).

Trotz des umfangreichen Angebots an »Einführungen in die Statistik für Sozialwissenschaftler« einen weiteren Einführungstext vorzulegen, scheint auf den ersten Blick überflüssig. Auf dem deutschen Buchmarkt ist jedoch nach unserer Kenntnis neben Armingers sehr guter, aber auch sehr gedrängter Darstellung (in Jarausch/Arminger/Thaller 1985) keine systematische Einführung in die statistische Methodenlehre erschienen, die speziell für Historiker verfaßt worden wäre. Die Erfahrung zeigt, daß sich Lernbarrieren im Zugang zu formalen Methoden am leichtesten überwinden lassen, wenn die entsprechenden Konzepte und Analyseverfahren anhand von Daten und Fragestellungen aus dem eigenen Fachgebiet vermittelt werden. Zwar sind einige Vielzweckbände über »Quantitative Methoden für Historiker« (oder ähnliche Titel) erschienen; in ihnen werden aber die eigentlich »statistischen« Methoden und Modellkonstruktionen viel zu knapp abgehandelt.

Statistik wird in diesem Skript als ein Instrumentarium präsentiert, mit dem Historiker (wie auch Wissenschaftler aus anderen Disziplinen) Daten analysieren, Informationen verdichten, Zusammenhänge und Strukturen in ihnen erkennen und, auf dieser Grundlage, theoretische Hypothesen explorieren oder testen können. Großer Nachdruck wird darauf gelegt, die Verknüpfung inhaltlicher (substanzwissenschaftlicher) Konzepte mit formal-statistischen Modellvorstellungen zu erörtern und die Anwendungsvoraussetzungen einzelner Verfahren zu klären.

Bei den Kursteilnehmern werden keine statistischen oder mathematischen Kenntnisse vorausgesetzt, die über das Niveau allgemeiner Schulbildung hinausgehen. Auswahl und Darstellung der verschiedenen Themen orientieren sich an folgenden Kriterien:

(1) Der angebotene »Stoff« sollte in einem Anfängerkurs innerhalb von 12 Doppelstunden zu bearbeiten sein. (Abschnitte, die bei einem ersten Lektüredurchgang ausgelassen werden können, sind mit einem Sternchensymbol (*) gekennzeichnet.) (2) Das Skript soll sich auch für das individuelle Studium außerhalb von Lehrveranstaltungen eignen. (3) Es soll den Zugang zur weiterführenden Literatur und zu komplexeren Verfahren (auf die fortlaufend hingewiesen wird) erleichtern (denn in der Forschungspraxis reichen die elementaren Verfahren, auch die in Teil II behandelten, häufig nicht aus).

Die Analysebeispiele sind mit dem Programmsystem SPSS^{x} (Statistical Package for the Social Sciences), Version 3.1, ausgeführt worden. Die entsprechenden »Befehle« für die an Großrechnern installierte Fassung werden im Text zitiert. Das Skript ist jedoch bewußt auf die Erörterung statistischer Konzepte und Verfahren begrenzt; es bietet keine Einführung in die elektronische Datenverarbeitung.

Zur Durchführung der Analysen standen vier Datensätze zur Verfügung: (1) Die Abgeordneten der Frankfurter Nationalversammlung 1848/49 (von Heinrich Best). (2) Die Abgeordneten der Reichstage von 1867 - 1918 (Heinrich Best). (3) Biographische Daten der SPD-Reichstagskandidaten 1898 - 1912 (Wilhelm H. Schröder). (4) Wahlkreisdaten zu den Reichstagswahlen von 1898 - 1912 (Wilhelm H. Schröder).

Ich danke den Mitarbeitern des ZHSF sowie Heinrich Best, Jörg Blasius, Steffen Kühnel, Herbert Odenthal und Kurt Sombert, die Teile des Skripts gelesen und mit hilfreichen Kommentaren versehen haben. Ralph Ponemereo und Kurt Sombert danke ich außerdem für ihre großzügige Unterstützung bei den EDV-Arbeiten.

Köln, Mai 1989 Helmut Thome

KAPITEL 6

Wahrscheinlichkeitstheoretische Grundlagen der induktiven Statistik

Warum muß sich der Sozialhistoriker überhaupt um Wahrscheinlichkeitstheorie kümmern? Darauf gibt es mehrere Antworten. Obwohl wir sie erst im Verlauf der weiteren Darstellung ausführlich erörtern werden, wollen wir vorab schon ein paar Hinweise geben: Der Sozialhistoriker untersucht nicht nur Einzelfälle und kleine Kollektive, sondern typischerweise Massenphänomene, die er bei einer großen Zahl von Personen beobachtet. Man denke z. B. an die Millionen Befragten, über die Zensusdaten vorliegen. Hier ist es oft unmöglich, zumindest unnötig, eine Totalerhebung vorzunehmen und die Zensusbögen aller Personen oder Haushalte auszuwerten. Für ein bestimmtes Forschungsvorhaben genügt in der Regel eine »repräsentative« Auswahl, eine »Stichprobe«, die vielleicht zwei- oder dreitausend Fälle umfaßt. Auf welcher Grundlage kann man die Repräsentativität einer solchen Auswahl einschätzen, und wie kann man auf der Basis der Stichprobenergebnisse verallgemeinernde Aussagen über die gesamte Population, die »Grundgesamtheit« machen? Die Mittel hierzu liefert die Wahrscheinlichkeitstheorie.

Wir benötigen die Wahrscheinlichkeitstheorie aber nicht nur in diesem Zusammenhang. Ein weiterer Anwendungsfall ergibt sich bei der Konstruktion probabilistischer Modelle, zum Beispiel bei der Zeitreihenanalyse: Der Forscher versucht anhand seiner Daten einen »Erzeugungsmechanismus« zu entdecken (oder ihn aus einer bereits vorgeschlagenen Theorie abzuleiten), der die beobachteten Erscheinungen »hervorgebracht« haben könnte. Dabei rechnet man auch für den Fall eines adäquaten Modells damit, daß die Beobachtungen aufgrund nie vollständig erfaßbarer Einflußfaktoren (zu denen auch Meßfehler gehören können) innerhalb eines Wertebereichs mit unterschiedlichen Wahrscheinlichkeiten schwanken. Die Abweichungen vom »Mittel« und ihre »Wahrscheinlichkeitsverteilungen« versucht man von vornherein in dem Modell zu berücksichtigen. So betrachtet man im Prinzip alle Beobachtungsdaten als Ergebnis eines »Zufallsexperiments«, auch wenn die Daten nicht aus einer Stichprobenziehung im üblichen Sinne hervorgegangen sind. Das theoretische Modell bezieht sich dann auf eine hypothetische (»konzeptuelle«), nicht auf eine empirische Population. Solche Überlegungen spielen natürlich nicht nur bei der Zeitreihenanalyse eine Rolle. Wir haben schon in Teil I

(siehe den Schluß der Abschnitte 5.2.5 und 5.3) angedeutet, wie auch »einfache« Modelle über den Zusammenhang oder die Unabhängigkeit zweier Variablen unter dieser Perspektive betrachtet werden können: wie gut »paßt« ein theoretisches Modell zu den empirischen Daten, wenn man Zufallseinflüsse mit in Rechnung stellt. Die in Kap. 10 erläuterte Regressionsanalyse wird diesen Gedankengang weiter verdeutlichen.

Wenn wir uns nun einige wahrscheinlichkeitstheoretische Konzeptionen aneignen wollen, müssen wir von sehr abstrakten Modellvorstellungen ausgehen, die zunächst kaum einen Anwendungsbezug zur Geschichtswissenschaft zu haben scheinen. Der Leser braucht also etwas Geduld.

6.1 Zufallsexperiment und Zufallsvariable

In seinen empirischen Erhebungen untersucht der Sozialwissenschaftler u.a. Häufigkeitsverteilungen und »Zusammenhänge« zwischen bestimmten Variablen (siehe Teil I dieses Skripts). Wo er Merkmalsausprägungen »mißt«, sie u. U. auch experimentell hervorbringt, spricht der Statistiker von »Ereignissen«, die er als Resultat von »Zufallsvorgängen« (Zufallsexperimenten, Zufallsmechanismen) interpretiert. Der Begriff **Zufallsexperiment** meint hier nicht eine bestimmte Methode der Datengewinnung, sondern ist allgemeiner gefaßt. Der Statistiker versteht darunter eine Aktion oder einen Vorgang, der im Prinzip unter den gleichen Bedingungen beliebig oft wiederholbar ist und dessen Ausgang auch nach wiederholten Ausführungen unsicher, nicht exakt vorhersagbar ist. »Unter gleichen Bedingungen wiederholbar« schließt die Voraussetzung ein, daß das eine Experiment keinerlei Einfluß auf die Ergebnisse des anderen Experiments hat.

Das Zufallsexperiment kann real oder gedanklich durchgeführt werden. Oft zitierte Beispiele sind das Werfen einer Münze oder eines Würfels oder das Ziehen einer Kugel aus einer Urne; aber auch das Befragen einer Person nach ihrer Parteipräferenz oder eines Zensusdokuments nach dem dort festgehaltenen Datum der Eheschließung einer bestimmten Person kann als ein Zufallsexperiment konzipiert werden, wenn das Ergebnis der Befragung nicht schon vorher feststeht. Es wird nicht verlangt, daß der Forscher den Zufallsvorgang selbst initiiert; er kann in der Rolle des passiven Beobachters bleiben. Er registriert dann gleichsam die Ergebnisse der Zufallsexperimente, die in der Natur oder der Gesellschaft ablaufen.

In dieser abstrakten Konzeption beinhaltet der Begriff des Zufallsexperiments keinerlei Annahmen über die kausale oder a-kausale Struktur des beobachteten Ereignisses. Es bleibt völlig offen, welche Bedingungen und Ursachen zu seinem Zustandekommen beigetragen haben. Er kenn-

zeichnet lediglich den Tatbestand, daß der Forscher nicht im voraus (nicht bevor er gemessen hat) das Resultat seiner Beobachtung kennt. Die unterschiedlichen »**Realisationen**«, die unterschiedlichen Ergebnisse des Zufallsvorgangs, sind, aus dieser Perspektive betrachtet, lediglich mehr oder weniger »wahrscheinlich«.

Als Zufallsexperiment läßt sich sowohl der einmalige Versuch (z. B. Befragen einer Person) als auch eine Gesamtheit wiederholter Versuche (z. B. Befragen von n Personen) betrachten. Dementsprechend ändert sich auch die Menge der möglichen Ergebnisse oder Ereignisse des Zufallsexperiments. Einelementige Ereignisse nennt man **Elementarereignisse**. Bei dem Zufallsvorgang »einen sechsflächigen Würfel werfen« gibt es die Elementarereignisse $E_1 = 1$, $E_2 = 2$, ... , $E_6 = 6$. Die Menge aller möglichen Elementarereignisse eines Zufallsvorgangs bezeichnet der Statistiker als **Stichprobenraum** (»sample space«) oder **Ereignisraum**, der mit dem griechischen Buchstaben Ω (Omega) symbolisiert wird. Ereignisse können somit als Unter- bzw. Teilmenge des Stichprobenraums definiert werden. Zusammengesetzte Ereignisse sind Aggregate von Elementarereignissen. Das zusammengesetzte Ereignis »eine gerade Zahl werfen« umfaßt die Menge $E = (2, 4, 6)$ als Untermenge des Stichprobenraums $\Omega = (1, 2, 3, 4, 5, 6)$.

Der Stichprobenraum und damit die Menge der Elementarereignisse ergibt sich aus der Definition des Zufallsexperiments bzw. der Zufallsvariablen. Der Zufallsvorgang »Geburt des ersten Kindes« mit der Ereignisdimension (der »Zufallsvariablen«, siehe unten) »Geschlecht des Kindes« hat zwei Elementarereignisse:

a) weiblich (w)
b) männlich (m)

Der Zufallsvorgang »Geburt dreier Kinder« mit der Zufallsvariablen »Geschlechterabfolge« hat folgende 8 Elementarereignisse:

(www)	(mmw)
(mww)	(mww)
(wmw)	(wmw)
(wwm)	(wwm)

Das zusammengesetzte Ereignis »Unter den drei neugeborenen Kindern sind zwei Jungen« umfaßt die drei Elementarereignisse

(wmm) (mwm) (mmw).

Der Begriff der Elementarereignisse ist also etwas unscharf, da es von der Definition bzw. dem »Modell« des Zufallsexperiments abhängt, was als einzelnes Element gelten soll. Gelegentlich wird auch eine andere Terminologie benutzt, indem man formuliert: »Das **Ereignis** A:= 'Unter den 3

Neugeborenen sind 2 Jungen' besteht aus 3 **Ergebnissen**.« In dieser Terminologie umfaßt der Stichprobenraum die Menge aller möglichen »Ergebnisse« eines bestimmten Zufallsvorgangs.

Für ein Zufallsexperiment lassen sich oft alternative Modelle konstruieren (siehe Kregel 1988, S. 2-5), wie z. B. für das Zufallsexperiment »Werfen zweier sechsflächiger Würfel« mit der Zufallsvariablen (siehe unten) »Summe der Augenzahlen«. Im ersten Modell definieren wir die (Elementar-) Ergebnisse als die Paare (i,k) der im ersten (i) und im zweiten (k) Wurf erzielten Augenzahlen, also z. B. (1,1) (1,2) ... (6,5) (6,6). Der Stichprobenraum aller möglichen Ergebnisse hat folglich 36 Elemente: $\Omega = \{(i,k): 1 \leqq i, k \leqq 6\}$. Eine bestimmte Untermenge daraus bildet das Ereignis A:= 'Die Summe der Augenzahlen ist mindestens 11, nämlich $A = \{(6,6)\ (6,5)\ (5,6)\}$. In einem zweiten Modell könnte man die Menge $\Omega = \{2, 3, \ldots, 11, 12\}$ als Stichprobenraum definieren, denn die natürlichen Zahlen 2 bis 12 sind die möglichen Augensummen nach denen gefragt wird. Das erste Modell ist dem zweiten vorzuziehen, weil es informationshaltiger ist und weil sich in ihm, wie wir noch sehen werden, die »Wahrscheinlichkeiten« für die verschiedenen Ergebnisse und Ereignisse leichter angeben lassen.

Das Bestreben des Forschers ist es also, etwas über die Wahrscheinlichkeiten ausfindig zu machen, mit denen bestimmte Ereignisse auftreten, d. h., er will den Zufallsvorgang nicht nur durch die Menge seiner Elementarereignisse beschreiben, sondern jedem Ereignis eine Zahl zuordnen, die die Wahrscheinlichkeit des Eintretens dieses Ereignisses (bei Ablauf des Zufallsexperiments) angibt. Diese Zuordnung geschieht auf der Basis bestimmter Regeln, den Axiomen und Lehrsätzen der Wahrscheinlichkeitstheorie, von denen wir im folgenden einige kennenlernen werden.

Wenn wir unsere Daten »statistisch« auswerten wollen, betrachten wir jede erhobene Merkmalsdimension als »Zufallsvariable« (groß) X und ihre beobachteten Ausprägungen (klein) x_i als Realisationen eines Zufallsexperiments im obigen Sinne. (Der Index $i = 1, 2, \ldots$ bezieht sich auf die durchnumerierten möglichen Ausprägungen bzw. Werte). Eine **Zufallsvariable** läßt sich allgemein definieren als »eine numerische Größe, deren Wert durch ein Zufallsexperiment bestimmt wird. Die Ergebnisse des Experiments sind entweder selbst numerisch oder es wird ihnen eine Zahl zugeordnet« (FU-Autorenkollektiv 1976, S.88b). Durch eine solche Zuordnungsvorschrift wird die Möglichkeit der Beschreibung von Zufallsexperimenten wesentlich verbessert. So wird es z. B. grundsätzlich möglich, die Wahrscheinlichkeiten mit Hilfe mathematischer Funktionen den einzelnen Merkmalsausprägungen (oder Klassen von Merkmalsausprägungen) zuzuordnen. Das wird später noch verdeutlicht. Jetzt ist es zunächst einmal erforderlich, den Begriff der Wahrscheinlichkeit näher zu bestimmen:

6.2 Zum Wahrscheinlichkeitsbegriff

In der »klassischen« Konzeption von **Laplace** wird die Wahrscheinlichkeit für das Ereignis A definiert als

$$(6\text{-}1) \qquad P(A) = \frac{f(A)}{n}$$

f(A) := Häufigkeit des Vorkommens von Ereignis A bei n gleichmöglichen Ereignissen.

Die Laplace-Wahrscheinlichkeit ist also identisch mit der relativen Häufigkeit, mit der das gerade interessierende, das im Sinne der Fragestellung »günstige« Ereignis im Stichprobenraum vorkommt. Die Definitionsgleichung läßt sich in Langschrift somit wie folgt schreiben:

$$(6\text{-}1') \qquad P(A) = \frac{\text{Anzahl der für A günstigen Elementarereignisse}}{\text{Anzahl aller gleichmöglichen Elementarereignisse}}$$

Beispiel: Der Stichprobenraum bestehe aus $n = 100$ Personen A_i ($i = 1,2, \ldots, 100$). Das interessierende Ereignis sei die zufällige Auswahl der 37. Person, A_{37}. Da jede Person nur einmal vorkommt, ist $f(A_{37}) = 1 = f(A_i)$ und $P(A_{37}) = 1/100 = P(A_i)$. Es sei nun ein zweites Zufallsexperiment definiert, das mit dem ersten technisch verknüpft ist: Beobachten der Parteipräferenz einer Person, die aus einer Menge von 100 Personen zufällig ausgewählt wird. Den Stichprobenraum bilden nun die Parteipräferenzen der 100 Personen, nicht die Personen als solche. Man muß also zwischen den Elementareignissen und ihren (physischen) »Trägern« (hier Personen) unterscheiden. Das interessierende Ereignis A sei die Parteipräferenz für die CDU (bei Auswahl einer Person). Nehmen wir an, eine Parteipräferenz für die CDU komme im Stichprobenraum zweiundvierzigmal vor. Dann ist gemäß (6-1) die Wahrscheinlichkeit $P(A) = P(CDU) = 42/100 = 0{,}42$.

Für die Stichprobenräume beider Beispiele gilt: Jedes Elementarereignis des Stichprobenraumes hat die gleiche Chance, ausgewählt zu werden. Die Definition der Laplace-Wahrscheinlichkeit beruht auf diesem »Gleichmöglichkeitsmodell«; sie ist nur anwendbar, wenn die Elementarereignisse vollständig abzählbar sind.

Nicht immer ist das Abzählen so einfach wie in den vorangegangenen Beispielen. Es wird schon schwieriger, wenn wir das Zufallsexperiment nicht als »zufällige Auswahl einer Person«, sondern als »zufällige Auswahl dreier Personen aus einer Menge von 100 Personen« definieren und

als interessierendes Ereignis A das dreimalige Auftreten der CDU-Präferenz bei den ausgewählten Personen betrachten. Die Elementarereignisse des Stichprobenraums stellen nun alle möglichen Dreierkombinationen dar, die sich aus den 100 Personen auswählen lassen. Im Vorgriff auf den Exkurs in Abschnitt 6.4 läßt sich ihre Zahl mit $n = 100!/(3!97!) = 161700$ angeben. Soviele Möglichkeiten gibt es, aus 100 Elementen drei auszuwählen, wenn deren Reihenfolge keine Rolle spielt und jedes Element nur einmal augewählt werden darf. Wenn sich unter den 100 Personen wiederum 42 CDU-Anhänger befinden, läßt sich die Zahl der interessierenden (»günstigen«) Elementarereignisse mit $f(A) = 42!/(3!39!) = 11480$ berechnen. Somit ist $P(A) = 11480/161700 = 0{,}071$.

Allerdings gibt es in diesem Beispiel noch eine andere Möglichkeit, P(A) zu betimmen, die vielleicht intuitiv zugänglicher ist als die erste: Das Ereignis A (Vorkommen dreier CDU Präferenzen bei dreimaliger Zufallsauswahl) läßt sich auch als Verknüpfung $E = A_1 + A_2 + A_3$ dreier Elementarereignisse auffassen. Die Wahrscheinlichkeit $P(A) = P(E)$ wird dann aus den Einzelwahrscheinlichkeiten $P(A_1)$, $P(A_2)$ und $P(A_3)$ errechnet - nach Regeln, die in Abschnitt 6.3 erläutert werden: $P(E) = P(A_1) \cdot P(A_2) \cdot (A_3) = [(42/100)(41/99)(40/98)] = 0{,}071$.

Die klassische Definition der Wahrscheinlichkeit ist offenkundig nur auf solche Zufallsexperimente anwendbar, bei denen der Stichprobenraum endlich viele, gleichwahrscheinliche Elementarereignisse enthält, deren Menge bekannt ist. Das bedeutet für die praktische Forschungssituation in vielen Fällen eine zu große Einschränkung. Häufig befinden wir uns ja gerade nicht in der Lage, die Anzahl der jeweils interessierenden oder im Stichprobenraum überhaupt befindlichen Elementarereignisse zu kennen, sondern sie aus Stichprobenergebnissen allererst erschließen zu wollen (wie z. B. die Häufigkeitsverteilung der Parteipräferenzen in der erwachsenen Bevölkerung, aus der lediglich eine Auswahl von 1000 Personen befragt wurde). Hier hilft der sog. **statistische Wahrscheinlichkeitsbegriff** weiter, den Richard von Mises 1918 vorschlug. Auch dieser Wahrscheinlichkeitsbegriff weist, sagen Mathematiker und Wissenschaftstheoretiker, Mängel auf. Aber er liefert ein Verfahren, mit dem Wahrscheinlichkeiten für Ereignisse bei den **praktisch** wichtigsten Zufallsexperimenten festgelegt werden können. Seine Definition bezieht sich wiederum auf die relative Häufigkeit, verbindet sie aber mit der Vorstellung eines **unendlich oft** wiederholten Zufallsexperiments. Die Wahrscheinlichkeit des Eintreffens eines bestimmten Ereignisses A wird definiert als Grenzwert seiner relativen Häufigkeit bei einer gegen unendlich strebenden Zahl von Versuchen:

$$(6\text{-}2) \qquad P(A) : \lim_{n\to\infty} \frac{f(A)}{n}$$

In dieser Formel steht »n« für die Zahl der Zufallsvorgänge (den Umfang der Stichprobe) und »f(A)« für die Häufigkeit, mit der das Ereignis A dabei auftritt. Was mit der Definitionsgleichung gemeint ist, läßt sich am Beispiel des Münzwurfs veranschaulichen. Wenn wir eine (faire) Münze n-mal werfen, liegt das Wappen a-mal und die Zahl (n-a)mal oben. Wir können dann eine zweite Versuchsreihe mit größerem n starten und danach weitere Versuchsreihen mit immer größer werdendem n. Die relative Häufigkeit, mit der »Wappen« auftritt, wird schließlich sehr nahe um den Wert 0,5 schwanken. Das bedeutet, daß uns die tabellarisch oder graphisch festgehaltenen Untersuchungsergebnisse (siehe Abb. 6.1) einen »guten Grund« liefern anzunehmen, daß der Grenzwert und damit die Wahrscheinlichkeit P(Wappen) = 0,5 ist.
Dabei können wir uns auf das sog. Bernoulli-Theorem, das (schwache) Gesetz der großen Zahl, berufen. Es besagt, daß mit einer wachsenden Zahl von Beobachtungen (Zufallsexperimenten) n die absolute Differenz zwischen der relativen Häufigkeit f (A)/n und der Wahrscheinlichkeit P(A) kleiner wird als eine vorgegebene, beliebig kleine positive Zahl; die Abweichungen der relativen Häufigkeit von der Wahrscheinlichkeit werden immer geringer. Voraussetzung für die Gültigkeit dieses Theorem ist, daß es sich wirklich um Zufallsexperimente handelt, die einzelnen Versuchsergebnisse also unabhängig voneinander realisiert werden. (Den Begriff der Unabhängigkeit werden wir gleich noch näher erörtern.)

In der Praxis läßt sich ein Experiment natürlich nicht unendlich oft wiederholen. Wahrscheinlichkeitsannahmen sind also letztlich Hypothesen, die gleichwohl mit der beobachtbaren Wirklichkeit mehr oder weniger gut vereinbar sind. Oft stützt man sich, wie bei den Beispielen des Werfens einer Münze oder eines Würfels, auf Plausibilitätsbetrachtungen oder, anders ausgedrückt, auf bestimmte Modellvorstellungen über den Mechanismus des Zufallsexperiments.

Auf weitere Wahrscheinlichkeitsbegriffe wie den der »subjektiven« Wahrscheinlichkeit gehen wir hier nicht ein.

6.3 Das Rechnen mit Wahrscheinlichkeiten

Wenn der Sozialwissenschaftler die empirische Grundgesamtheit, aus der er eine Stichprobe ziehen will, kennt, wenn er über die in ihr gegebenen Häufigkeitsverteilungen Bescheid weiß, kann er (gemäß (6-1) diese relativen Häufigkeiten unmittelbar als Wahrscheinlichkeitsverteilung für jeden einzelnen Versuch bei der Stichprobenziehung interpretieren. Wenn er z. B. weiß, daß 10 % aller Reichstagsabgeordneten zwischen 1871 und 1918 unter 40 Jahre alt waren, so ergibt sich bei jeder zufälligen Auswahl aus dieser Grundgesamtheit eine Wahrscheinlichkeit von $p = 0{,}10$ dafür, daß man einen unter vierzigjährigen Abgeordneten »herauszieht«. (Das Konzept der Zufallsauswahl wird in Kap. 9 näher erläutert.) In vielen Forschungssituationen sind aber die Häufigkeitsverteilungen in der Grundgesamtheit, der »Population« (also der Menge aller Fälle, aus der man eine Stichprobe zieht und über die man letztlich eine Aussage machen will) gar nicht bekannt. Wenn man keine Totalerhebung vornehmen kann oder will, ist man darauf angewiesen, aus den Ergebnissen von Zufallsexperimenten (also von Stichprobenergebnissen) auf bestimmte Kennwerte der Populationsverteilung (z. B. den Mittelwert) zu schließen, solche Kennwerte also zu schätzen. Dann interpretiert man die Kennwerte als Zufallsvariablen, die mit unterschiedlichen Wahrscheinlichkeiten verschiedene Werte annehmen können. Das werden wir in Kap. 7 näher besprechen. Aber schon jetzt dürfte deutlich sein: Wenn wir jemals Wahrscheinlichkeiten für solche spezielle Zufallsvariablen »ableiten« wollen, müssen wir allgemein wissen, wie man aus bekannten Wahrscheinlichkeiten für Elementarereignisse Wahrscheinlichkeiten für Ereignisverknüpfungen errechnet (das arithmetische Mittel zum Beispiel ergibt sich aus der Verknüpfung einer Reihe von Einzelbeobachtungen). Um diese Frage beantworten zu können, müssen wir zunächst ein paar terminologische Konventionen einführen (siehe Hartung/Elpelt/Klösener 1986, S. 92):

(6-3) $(A \cup B)$ sei das Ereignis, daß **mindestens** eines der beiden Ereignisse A und B auftritt. (Man spricht auch von der »Vereinigungsmenge«)

(6-4) $(A \cap B)$ sei das Ereignis, daß **sowohl** A **als auch** B eintreffen. Die Menge dieser Ereignisse (die man auch als »Schnittmenge« bezeichnet), ergibt die sog. Nullmenge, wenn A und B **disjunkte**, d.h. wechselseitig sich ausschließende Ereignisse sind, sie also gar nicht gemeinsam auftreten können.

(6-5) $A - B$ sei das Ereignis, daß A, nicht aber B eintritt.

(6-6) $A = \Omega - A$ soll bedeuten, daß A **nicht** eintritt, und heißt »das zu A komplementäre Ereignis«.

Zur Veranschaulichung dieser Definitionen bedient man sich der sog. Mengendiagramme, in denen unterschiedliche Typen von (Teil-)Mengen-Beziehungen dargestellt sind (siehe Abb. 6.2).

Um Wahrscheinlichkeiten für verknüpfte Ereignisse berechnen zu können, stützt man sich auf folgende Axiome (6-7) bis (6-9):

(6-7) $P(A) \geq 0$

(6-8) $P(\Omega) = 1$

Diese beiden Axiome bedeuten vor allem eine Normierung: Man legt die Wahrscheinlichkeiten so fest, daß sie nicht kleiner als »0« und nicht größer als »1« (sog. sicheres Ereignis) sind. Außerdem müssen sich die Wahrscheinlichkeiten für alle möglichen Ereignisse eines Zufallsexperiments in der Summe zu »1« addieren.

(6-9) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ für jede Folge paarweiser disjunkter Ereignisse A_i

Dieses dritte Axiom beinhaltet auch die »endliche Additivität«, das heißt, für disjunkte Ereignisse $A_1, \ldots, A_n$ gilt:

(6-9′) $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Bei zwei disjunkten Ereignissen gilt also:

(6-9′′) $P(A \cup B) = P(A) + P(B)$

Dieses Axiom wird in der Literatur häufig als **»Additionstheorem«** eingeführt. Wir werden von ihm im nächsten Kapitel noch Gebrauch machen. Hier nur ein Beispiel aus dem Würfelmodell: Wie groß ist die Wahrscheinlichkeit P(A U B), daß eine Zwei (Ereignis A) oder eine Fünf (Ereignis B) geworfen wird? Es ist $P(A) = 1/6$, $P(B) = 1/6$, folglich $P(A \cup B) = P(A) + P(B) = 2/6$.

Wie gesagt, das Additionstheorem gilt in dieser Form nur für disjunkte (innerhalb eines Zufallsvorganges sich wechselseitig ausschließende) Ereignisse (bei einem Wurf z. B. können nicht sowohl die Zwei als auch die Fünf vorkommen). Für zwei **beliebige** (evtl. nicht-disjunkte) Ereignisse lautet die Verknüpfungsregel:

(6-10) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Der negative Summand für die Schnittmenge wird Null, wenn die Ereignisse A und B sich wechselseitig ausschließen. Warum man die Wahrscheinlichkeit für die Schnittmenge der beiden Ereignisse abziehen muß, kann man sich anhand der Mengendiagramme in Abb. 6.2 verdeutlichen: Die Schnittmenge von A und B wird ja zunächst sowohl bei P(A) als auch bei P(B), also zweimal berücksichtigt. Folglich muß sie einmal wieder abgezogen werden.

Zwei weitere Rechenregeln, die sich (ohne daß wir das hier zeigen) aus den Axiomen (6-7) bis (6-9) ableiten lassen, seien noch erwähnt:

(6-11) $P(\bar{A}) = 1 - P(A)$

(6-12) $P(A-B) = P(A) - P(B)$, falls B eine Teilmenge von A ist.

Der oben benutzte Begriff der disjunkten Ereignisse ist scharf von dem Begriff der (stochastisch) **unabhängigen Ereignisse** zu trennen. Bevor wir uns klar machen, was darunter zu verstehen ist, müssen wir noch einen weiteren Begriff einführen, nämlich den der **bedingten Wahrscheinlichkeit**. Dieses Konzept läßt sich wiederum in Analogie zu den bedingten Häufigkeiten definieren, die wir in der Tabellenanalyse kennengelernt haben. Zu diesem Zweck drucken wir hier noch einmal die Tabelle aus Abb. 4.2 (Teil I):
Stellen wir fest, wie häufig das Merkmal y_2 (mittlere Mobilität) unter der Bedingung, daß x_1 (niedrige Schulbildung) vorliegt, vorkommt:

$$f(y_2 | x_1) = 38$$

Die Bedingung x_1 ist insgesamt 88mal gegeben. Folglich ist die bedingte relative Häufigkeit

$$f_r(y_2 | x_1) = 38/88 = = 0{,}4318$$

Analog definieren wir nun die **bedingte Wahrscheinlichkeit** eines Ereignisses A (»y_2 ist realisiert«) unter der Bedingung B (»x_1 ist realisiert«) allgemein für $P(B) > 0$ als

(6-13) $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{f(A \cap B) / n}{f(B) / n} = \frac{f(A \cap B)}{f(B)}$$

In unserem Beispiel: (38/417)/(88/417) = 33/88

In Fällen wie diesem läßt sich die Häufigkeit, mit der die Ereignisver-

knüpfung ($A \cap B$) vorkommt, direkt im Sinne des Gleichmöglichkeitsmodells auszählen. Eine Regel, wie sich die Wahrscheinlichkeit für dieses zusammengesetzte Ereignis aus den beiden Wahrscheinlichkeiten P(A) und P(B) errechnen läßt, wenn A und B disjunkte Ereignisse sind, wird weiter unten erläutert.

Nun können wir auch den Begriff der **stochastischen Unabhängigkeit** definieren:

Die Ereignisse A und B sind (stochastisch) unabhängig voneinander, wenn $P(B|A) = P(B|\overline{A})$, wenn also die Wahrscheinlichkeit, daß B unter der Bedingung A eintritt, genauso groß ist wie unter der Bedingung, daß A nicht eintritt. In diesem Falle sind die bedingten Wahrscheinlichkeiten $P(B|A) = P(B)$ und $P(A|B) = P(A)$. Disjunkte Ereignisse sind gerade **nicht** unabhängig voneinander, da ja B nicht auftreten kann, wenn A auftritt und umgekehrt, in diesem Falle ist also $P(A|B) = P(B|A) = 0$.

Stochastisch unabhängige Ereignisse sind z. B.:

A:= aus einem verdeckten, gut gemischten Skatkartensatz ein As ziehen.

B:= aus einem verdeckten, gut gemischten Skatkartensatz eine rote Karte ziehen

Es ist (nach dem Gleichmöglichkeitsmodell)

$P(A) = 4/32 = 2/16$, da es unter den 32 Karten, die den Stichprobenraum bilden, genau 4 Asse gibt.

$P(B) = 16/32 = 2/4$, da es unter den 32 Karten genau 16 rote Karten gibt.

$P(A|B) = 2/16$, da es unter 16 roten Karten (der Bedingung B) genau 2 Asse (Ereignis A) gibt.

$P(B|A) = 2/4$, da es unter vier Assen (der Bedingung A) genau 2 rote Karten (Ereignis B) gibt.

$P(A \cap B) = 2/32$, da es unter den 32 Karten genau 2 Karten gibt, die sowohl rot sind (Ereignis B) als auch ein As darstellen (Ereignis A).

Somit erhalten wir in diesem Falle folgende Beziehungen:

$P(A) = P(A|B) = 2/16$
$P(B) = P(B|A) = 2/4$
$P(A \cap B) = P(A)P(B) = (2/16)(1/2) = 2/32$

Diese Beziehungen gelten allgemein, wenn zwei (nicht-disjunkte) Ereignisse stochastisch unabhängig sind. Deshalb benutzt man

$$(6\text{-}14) \qquad P(A \cap B) = P(A)P(B)$$

auch als Definition der Unabhängigkeit zweier Ereignisse A und B. In der Literatur wird diese Beziehung als **»Multiplikationstheorem«** eingeführt; es läßt sich auf beliebig viele Ereigniswahrscheinlichkeiten ausdehnen.

Für den allgemeinen Fall, der die Möglichkeit einschließt, daß die Ereignisse nicht unabhängig sind, muß (6-14) modifiziert werden:

$$(6\text{-}15) \qquad P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

Der allgemeine **Multiplikationssatz für bedingte Wahrscheinlichkeiten** lautet:

$$(6\text{-}16) \qquad P(A_1 \cap \ldots \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdot\ldots \cdot P(A_k|A_1 \cap \ldots \cap A_{k-1})$$

Damit haben wir durchaus nicht alle für die Statistik relevanten wahrscheinlichkeitstheoretischen Sätze dargestellt. Wir werden jedoch bei den praktischen Beispielen dieses Kurses lediglich das Additionstheorem und das Multiplikationstheorem benutzen.

6.4 Exkurs: Permutationen und Kombinationen (*)

Das »Abzählen« von Ereignismöglichkeiten läßt sich erleichtern, wenn man einigen Regeln der Kombinatorik folgt. Sie geben u. a. an, wieviel Möglichkeiten es gibt, n Elemente anzuordnen (»Permutationen«) oder von n Elementen k Elemente auszuwählen (»Kombinationen«). (Diese Terminologie ist in den entsprechenden Lehrbüchern nicht einheitlich gehandhabt; wir folgen hier den Begriffsbestimmungen in Hartung/Elpelt/Klösener (1986), S. 96.)

Permutationen

Zwei Fälle sind zu unterscheiden:

a) Alle n Elemente sind verschieden; dann gibt es

$$(6\text{-}17) \qquad n! \text{ (lies "n Fakultät")} = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$$

Möglichkeiten der Anordnung.

b) Einige (viele) der n Elemente sind untereinander gleich, so daß man von s, $s < n$, unterschiedlichen Typen i ($i = 1, 2, \ldots, s$) mit jeweils n_i Elementen sprechen kann. In diesem Falle gibt es

$$(6\text{-}18) \qquad \frac{n!}{n_1!n_2!n_3!\ldots n_s!} \quad , \quad n_1+n_2+\ldots+n_s=n$$

unterschiedliche Möglichkeiten, die Elemente anzuordnen (wobei es keine Rolle spielt, in welcher Reihenfolge die Elemente innerhalb der Typen angeordnet sind, da sie ja laut Voraussetzung gleich sind).

Beispiel: Bei 10 nach ihrer Parteipräferenz befragten Personen gebe es 4 CDU-Anhänger, 3 SPD-Anhänger, 2 Grünen-Anhänger und 1 FDP-Anhänger. Wenn diese Personen nur hinsichtlich ihrer Parteineigung als gleich oder ungleich eingestuft werden, gibt es $(10!)/(4!3!2!1!) = 12600$ verschiedene Möglichkeiten, sie in einer Reihe anzuordnen. Falls man verlangt, daß die Anhänger ein und derselben Partei jeweils hintereinander aufgereiht werden, gibt es nur noch $4! = 24$ Anordnungsmöglichkeiten, da die 4 Typen nun jeweils einen festen Block bilden, sich also die effektive Zahl der Elemente von 10 auf 4 verringert (wenn die Reihenfolge der Personen innerhalb des Blocks gleichgültig ist).
Wenn es nur einen mehrelementigen Typus mit k, $k > 1$, gleichen Elementen gibt, während alle anderen n-k Elemente unterschiedlich sind, reduziert sich (6-18) zu

$$(6\text{-}18') \qquad \frac{n!}{k!} \qquad \text{Anordnungsmöglichkeiten.}$$

Wenn sich alle Elemente jeweils dem einen oder dem anderen von nur zwei vorhandenen Typen zuordnen lassen, ergibt sich für den zweiten Typ die Zahl der Elemente aus $n_2 = n - n_1$. Die Formel (6-18) reduziert sich dann zu

$$(6\text{-}18'') \qquad \frac{n!}{n_1!(n-n_1)!} = \binom{n}{n_1} \qquad \text{, lies "n über } n_1\text{"}$$

Wenn alle Elemente gleich sind, gibt es überhaupt keine Möglichkeit, die Anordnung zu variieren. Man definiert also: $\binom{n}{0} := 1$ und $0! := 1$.

Kombinationen

Hier geht es um die Frage, wieviele verschiedene Möglichkeiten es gibt, aus einer Menge von n **verschiedenen** Elementen k Elemente mit oder ohne Wiederholung auszuwählen. Es ist zu unterscheiden, ob die Reihenfolge der k Elemente berücksichtigt werden soll oder nicht:

a) Bei Kombinationen ohne Wiederholung und ohne Berücksichtigung der Reihenfolge gibt es

$$(6\text{-}19) \qquad \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Möglichkeiten der Anordnung. So hat man z. B. beim Sechserlotto $49!/(43!\cdot 6!) = 13983816$ Möglichkeiten, aus den 49 Zahlen sechs unterschiedliche auszuwählen. Wir erhalten hier bei veränderter Fragestellung das gleiche Ergebnis wie zu (6-18"). Berücksichtigte man auch noch die Reihenfolge, würde sich die Zahl der Möglichkeiten noch einmal drastisch erhöhen auf

$$(6\text{-}20) \qquad \frac{n!}{(n-k)!} = \frac{49!}{43!} = 10068347000$$

b) Läßt man zu, daß die Elemente auch mehrfach ausgewählt werden (»Kombinationen mit Wiederholung«), so gibt es bei unberücksichtigter Reihenfolge

$$(6\text{-}21) \qquad \frac{(n+k-1)!}{k!(n-1)!}$$

Möglichkeiten, dies zu tun. Da Wiederholungen zugelassen sind, kann die Zahl k der zu ordnenden Elemente größer n sein. Sind z. B. in einer größeren Gruppe alle Personen Mitglied in einer von vier überhaupt vertretenen Parteien (d. h., daß es 4 Arten von Elementen gibt), so gibt es $(4+7-1)!/(7!(4-1)!) = 120$ Möglichkeiten Untergruppen von je 7 Personen zu bilden, die sich hinsichtlich ihrer parteilichen Zusammensetzung unterscheiden (im Extremfall neigen alle sieben Personen zur gleichen Partei).
Wenn man dagegen die Reihenfolge der Personen mit berücksichtigt, wenn also die Untergruppen nicht nur danach unterschieden werden sollen, mit wievielen Anhängern jede Partei in ihr vertreten ist, dann gibt es $n^k = 4^7 = 16384$ Möglichkeiten der Anordnung.

Den Kombinationen »mit Wiederholung« entspricht in der Stichprobentheorie die Zufallsauswahl »mit Zurücklegen« des jeweils gezogenen Elements in die Grundgesamtheit, den Kombinationen »ohne Wiederholung« die Zufallsauswahl »ohne Zurücklegen« (siehe Kap. 9).

6.5 Wahrscheinlichkeitsverteilungen und ihre Kennwerte

So wie wir relative Häufigkeiten nicht nur als einzelne Größen, sondern in ihrer »Verteilung« über alle Ausprägungen einer Variablen betrachtet haben, können wir nun auch Verteilungen für Wahrscheinlichkeiten definieren. Dabei müssen wir den Fall diskreter Variablen von dem stetiger bzw. kontinuierlicher (als kontinuierlich behandelter) Zufallsvariablen unterscheiden.

Die formale Regel, nach der den einzelnen Merkmalsausprägungen, also den durch die diskrete Zufallsvariable X festgelegten möglichen Ereignissen, Wahrscheinlichkeiten zugeordnet werden, heißt **Wahrscheinlichkeitsfunktion** $f(x) = P(X = x) = p(x)$, wobei $\Sigma\, p_i\,(x_i) = 1$. Ihre graphische Darstellung ähnelt dem Histogramm bzw. Stabdiagramm für relative Häufigkeiten (siehe Abb. 6.4).

Für eine kontinuierliche Zufallsvariable ist keine »Wahrscheinlichkeitsfunktion« definiert, sondern eine sog. **Wahrscheinlichkeitsdichtefunktion.** Das Konzept der Dichtefunktion haben wir in bezug auf relative Häufigkeiten schon in Teil I, Abschnitt 2.4 erläutert. Den konzeptuellen Übergang von der relativen Häufigkeitsdichte bei klassierten Merkmalen zur **Wahrscheinlichkeit**sdichtefunktion können wir uns in einem Gedankenexperiment verdeutlichen, in dem wir die Zahl der Fälle immer weiter erhöhen und Klassenintervalle fortlaufend verkleinern. Im Grenzfall haben wir statt der einzelnen Säulen eine beliebig teilbare Fläche unter einer glatten Kurve, die die Dichtefunktion repräsentiert. Die folgende Abbildung, die wir dem Lehrbuch von Wonnacott/Wonnacott (1972) entnehmen, veranschaulicht die Abfolge dieser Schritte am Beispiel der Variable »Körpergröße von Männern« (s. Abb. 6.5).

Die Dichtefunktion, die man ebenfalls mit f(x) bezeichnet, ist also eine stetige, nicht-negative Funktion, wobei der Flächeninhalt unter der Kurve über einem bestimmten Intervall die Wahrscheinlichkeit angibt, daß die Zufallsvariable X einen Wert aus diesem Intervall annimmt. Anders als bei der Wahrscheinlichkeitsfunktion für diskrete Variablen ist der bestimmte Wert $f(x_i)$ an der Stelle x_i nicht direkt als Wahrscheinlichkeit interpretierbar. Nur das Flächensegment über einem Intervall [a,b], sein Anteil an der Gesamtfläche, ist als Wahrscheinlichkeit $P(a < x \leqq b)$ interpretierbar. Wahrscheinlichkeiten mit $P > 0$ können folglich nur für Intervalle, nicht für (ausdehnungslose) Punkte des Wertekontinuums angegeben werden. Es wäre ein nutzloses Unterfangen, wollte man die »Wahrscheinlichkeitsfunktion« für eine stetige Zufallsvariable ermitteln. Die Chance, einen Wert »exakt« zu treffen, ist gleich Null, wenn nach dem Komma des exakten Wertes beliebig viele Stellen folgen.

Bei der Erörterung der Häufigkeitsverteilung haben wir auch kumulierte Verteilungsformen besprochen. Kumulierte Verteilungen sind für Wahrscheinlichkeitsbetrachtungen besonders wichtig, da man häufig wissen möchte, wie groß die Wahrscheinlichkeit ist, daß ein bestimmter Wert der Zufallsvariablen nicht über- oder unterschritten wird. Man läßt in diesem Kontext häufig das Adjektiv »kumuliert« weg und spricht allgemein von der (Wahrscheinlichkeits-)**Verteilungsfunktion** $F(x) = P(X \leqq x)$. Die Verteilungsfunktion beantwortet also die Frage: Wie groß ist die Wahrscheinlichkeit, daß beim Zufallsvorgang ein Wert $X = x$ realisiert wird, der nicht oberhalb eines bestimmten Wertes x_m liegt.

Bei **diskreten** Zufallsvariablen (mit Rangordnung) erhält man diese Wahrscheinlichkeit, indem man die Wahrscheinlichkeiten aller x_i mit $i = 1,2, \ldots, m$ summiert:

$$(6\text{-}22) \qquad F(x_m) = \sum_{i=1}^{m} f(x_i)$$

Die Verteilungsfunktion einer diskreten Zufallsvariablen ist somit eine Treppenfunktion. Bei den Realisationsmöglichkeiten x_i weist sie Sprünge mit der Höhe $P(X = x_i)$ auf. Die Abbildung 6.6 zeigt die Verteilungsfunktion F(x), die sich aus der in Abb. 6.4 dargestellten Wahrscheinlichkeitsfunktion ergibt.

Bei **kontinuierlichen** Zufallsvariablen erhalten wir die Verteilungsfunktion als Integral aus der Dichtefunktion:

$$(6\text{-}23) \qquad F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

Die Wahrscheinlichkeit, daß eine Realisation ($x \leqq a$) eintritt, ist also

$$F(a) = P(X \leq a) = P(-\infty < x \leq a) = \int_{-\infty}^{a} f(x)\,dx.$$

Die Abbildung 6.7 stellt eine Dichtefunktion f(x) und die entsprechende Verteilungsfunktion F(x) einander gegenüber.

Umgekehrt erhält man die Dichtefunktion als 1. Ableitung aus der Verteilungsfunktion:

$$(6\text{-}24) \qquad f(x) = \frac{\partial F(x)}{\partial x}$$

Grundsätzlich lassen sich Verteilungsfunktionen mit zweierlei Absichten konstruieren: Man kann erstens eine empirische Häufigkeitsverteilung,

die man beobachtet hat, durch ein bestimmtes Verteilungsmodell »sparsam«, aber hinlänglich genau **beschreiben** wollen. Andererseits kann man auch aus theoretischen Annahmen ein **a-priori** Verteilungsmodell herleiten. Beispiele für den erstgenannten Zweck liefern u. a. die verschiedenen Modelle (z. B. »Pareto-Verteilung«), die das Charakteristische der Einkommensverteilung in einer bestimmten Nation festhalten. Als Beispiel für die **theoretische** Modellbestimmung ohne empirische Daten wollen wir hier eine Verteilungsfunktion für die Wartezeit bis zur Busankunft skizzieren (wir folgen hier Schlittgen 1987, S. 96 ff.)[1]:

Nehmen wir an, eine Person A gehe, ohne auf die Uhr zu schauen, zur Haltestelle und warte auf den nächsten Bus. Die Busse fahren alle 20 Minuten. Die stetige Zufallsvariable »X = Wartezeit in Minuten« kann dann jeden Wert aus dem Intervall [0;20] annehmen. Wir schließen Verspätungen und zu frühes Ankommen der Busse aus. Es ist also $P(0 \leqq X \leqq 20) = 1$ (»sicheres Ereignis«). Da die Versuchsperson ihre Abmarschzeit nicht mit dem Fahrplan koordiniert, ist es plausibel, in unserem Modell gleich langen Teilintervallen gleiche Wahrscheinlichkeiten zuzuordnen. Das bedeutet, daß die Dichtefunktion von X über dem Intervall [0;20] konstant ist: $f(x) = k$. Folglich ist die Wahrscheinlichkeit, eine Wartezeit zwischen zwei Zeitpunkten a und b, $0 < a < b < 20$, zu erhalten, proportional zu der Länge des Intervalls [a;b]: $P(a \leqq X \leqq b) = k(b-a)$. (Dabei ist k ein noch zu bestimmender Faktor.) Für ein beliebiges Ereignisintervall [a;b] erhalten wir die Wahrscheinlichkeit durch Integrieren:

$$(6\text{-}25) \qquad \int_a^b k(dx) = (k(b-a)$$

Jetzt müssen wir noch den Wert »k« bestimmen. Für $a = 0$ und $b = 20$, das »sichere Ereignis«, kennen wir die Wahrscheinlichkeit: $P(0 \leqq X \leqq 20) = k \cdot 20 = 1$. Daraus folgt $k = 1/20$. Somit lautet die Dichtefunktion von X:

$$(6\text{-}26) \qquad f(x) = \begin{cases} 1/20\ , & 0 \leq x \leq 20 \\ 0 & \text{sonst.} \end{cases}$$

Die Wahrscheinlichkeit, daß die Person A höchstens 5 Minuten wartet, beträgt also $(1/20)5 = 0{,}25$. (Da die Dichtefunktion eine Konstante darstellt, ist es gleichgültig, welche Werte wir für die Intervallgrenzen a und b

[1] Der Band von Schlittgen sei hier generell zur vertiefenden Lektüre empfohlen.

beim Integrieren einsetzen, solange b-a=5 ist). Die Wahrscheinlichkeit dafür, daß sie genau 5 Minuten wartet, ist dagegen gleich Null.

In der Regel braucht der Sozialwissenschaftler nicht selbst irgendwelche Verteilungsmodelle herzuleiten. In der einschlägigen Literatur findet er eine Vielzahl von Verteilungsmodellen, aus denen er das für seine Forschungszwecke geeignete »Exemplar« aussuchen kann. In diesem Grundkurs werden wir nur einige wenige Modelle kennenlernen, die für die Inferenzstatistik besonders wichtig sind, weil sie Verteilungen für gebräuchliche statistische Kennzahlen angeben.

Ebenso wie die (empirischen) Häufigkeitsverteilungen (siehe Teil I.,Kap. 3) lassen sich auch die (theoretischen) Wahrscheinlichkeits- bzw. Dichtefunktionen mit entsprechenden Kennwerten charakterisieren. Als Symbole verwendet man hierfür griechische Buchstaben; für das arithmetische Mittel z. B. den Buchstaben »μ«. Analog zur Gleichung (3-2) in Teil I. erhält man für diskrete Zufallsvariablen das arithmetische Mittel der Wahrscheinlichkeitsverteilung aus

$$(6\text{-}27) \qquad \mu = \sum_{i=1}^{n} x_i p_i$$

Bei kontinuierlichen Variabeln orientieren wir uns zwar auch an der Formel (3-2), müssen dabei aber gemäß Gleichung (2-6) die Klassenhäufigkeit durch das Produkt von Häufigkeitsdichte und Klassenbreite ersetzen und anschließend den Grenzübergang »Vergrößern des Stichprobenumfanges und Verkleinern der Klassenbreiten« (gedanklich) vollziehen. Somit erhalten wir statt der Summe das Integral

$$(6\text{-}28) \qquad \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

Man bezeichnet das arithmetische Mittel einer Wahrscheinlichkeitsverteilung auch als **Erwartungswert** E(X) der Zufallsvariable X oder der Verteilung von X. (Wir benutzen den Ausdruck »Wahrscheinlichkeitsverteilung« als Oberbegriff für Wahrscheinlichkeits- und Wahrscheinlichkeitsdichtefunktionen). Er läßt sich als derjenige Wert interpretieren, mit dessen Eintreffen (bei stetigen Verteilungen) vor der Durchführung des Zufallsexperiments am ehesten zu rechnen ist. Besonders anschaulich wird das bei der Lebenserwartung. Man kann z. B. fragen, welche weitere Lebenserwartung hat ein Mann, der gerade das 20. Lebensjahr vollendet hat und in der Bundesrepublik lebt. Aus der amtlichen Statistik, den sog. Sterbetafeln, lassen sich Wahrscheinlichkeitsverteilungen für die Zufallsvariable »X:= Lebensdauer in Jahren« oder für die Zufallsvariable »X:= Lebens-

dauer in Dekaden« berechnen. Im zweiten Falle bedeutet z. B. X = 1, daß der Mann eine Dekade überlebt, also das dreißigste Lebensjahr erreicht, aber vor Erreichen des vierzigsten stirbt. Wir entnehmen Schlittgen 1987, S. 124 die in Abb. 6.9 dargestellte Wahrscheinlichkeitsfunktion:
Daraus ergibt sich der Erwartungswert

$$(6\text{-}29)\quad E(X) = \sum_{i=1}^{9} x_i p_i = 0 \cdot 0{,}017 + 1 \cdot 0{,}021 + \ldots + 8 \cdot 0{,}001$$

$$= 4{,}705$$

Ein gerade 20 Jahre alt gewordener Mann kann demnach erwarten, weitere 4,705 Dekaden zu leben, also 67 Jahre alt zu werden.

Ein entsprechendes Beispiel für kontinuierliche Zufallsvariablen erhalten wir durch die oben (siehe Gleichung 6-26) konstruierte Dichtefunktion der Zufallsvariablen »X = Wartezeit für den Bus«.

Durch Anwendung von (6-28) ergibt sich die mittlere Wartezeit

$$(6\text{-}30)\quad E(X) = \int_0^{20} x \frac{1}{20}\, dx = \frac{1}{20}\left[\frac{1}{2}x^2\right]_0^{20} = \frac{1}{40}(20^2 - 0^2) = 10$$

Wer also ohne Blick auf den Fahrplan losmarschiert, muß mit einer (durchschnittlichen) Wartezeit von zehn Minuten rechnen.

Der Erwartungswert hat formale Eigenschaften, die denen des arithmetischen Mittels (siehe Teil I, S.36) analog sind. Es gilt also:

$$(6\text{-}31)\quad E(aX+b) = aE(X) + b \quad \text{für beliebige reelle Zahlen } a, b$$

$$E(X+Y) = E(X) + E(Y)$$

Das Rechnen mit Erwartungswerten (siehe Anhang) wird dadurch besonders ergiebig, daß jede stetige Funktion g(X) einer Zufallsvariablen X wieder als Zufallsvariable Y = g(X) aufgefaßt werden kann. Für diese neue Zufallsvariable lassen sich die Erwartungswerte analog zu (6-29) und (6-30) definieren:

$$(6\text{-}32)\quad E(Y) = E[g(x)] = \sum_i g(x_i)p_i \quad \text{für diskrete } X$$

$$(6\text{-}33) \quad E(Y) = E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad \text{für stetige X}$$

Auch die theoretische **Varianz** V(X) wird analog zur empirischen definiert (siehe Teil I, S.39). Als Symbol benutzt man jetzt ein kleines Sigma, σ^2:

$$(6\text{-}34) \quad V(X) = \sigma^2 = \Sigma\,(x_i - \mu)^2 p_i \quad \text{, wenn X diskret ist}$$

$$(6\text{-}35) \quad V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \quad \text{, wenn X stetig ist.}$$

Die Standardabweichung ist demgemäß $\sigma = \sqrt{\sigma^2}$.

Die Varianz läßt sich auch als Erwartungswert definieren, wenn wir $g(X) = (x - \mu)^2$ als spezielle Transformation der Zufallsvariablen X betrachten. Dann ist

$$(6\text{-}36) \quad E[g(X)] = E[(X-\mu)^2] = E[(X - E(X))^2]$$

Die Varianz V(X) ist also die erwartete quadratische Abweichung der Variablen X von ihrem Erwartungswert μ. Sie besitzt u. a. folgende Eigenschaften (zur Beweisführung siehe Anhang):

$$(6\text{-}37) \quad V(X) = E(X^2) - [E(X)]^2$$

Die Varianz V(Y) einer linear transformierten Zufallsvariablen $Y = a + bX$ ergibt sich aus

$$(6\text{-}38) \quad V(a+bX) = b^2V(X)$$

Wenn zwei Zufallsvariablen X und Y stochastisch unabhängig sind (siehe unten), ergibt sich die Varianz ihrer Summe aus der Summe der Einzelvarianzen:

$$(6\text{-}39) \quad V(X + Y) = V(X) + V(Y)$$

Wir haben in Teil I dieses Skripts nicht nur univariate, sondern auch zwei- und mehrvariate Häufigkeitsverteilungen kennengelernt, die wir in Tabellenform dargestellt haben. Auch für Wahrscheinlichkeiten können wir (wiederum unter Rückgriff auf das erwähnte Theorem von Bernoulli) in analoger Weise mehrdimensionale (»gemeinsame«) Verteilungen definieren, die u. a. durch »Randwahrscheinlichkeiten« und »bedingte Wahrscheinlichkeiten« definiert sind.

Wir hatten schon weiter oben die Unabhängigkeit zweier Ereignisse mit Hilfe des Begriffs der bedingten Wahrscheinlichkeit definiert (siehe Glei-

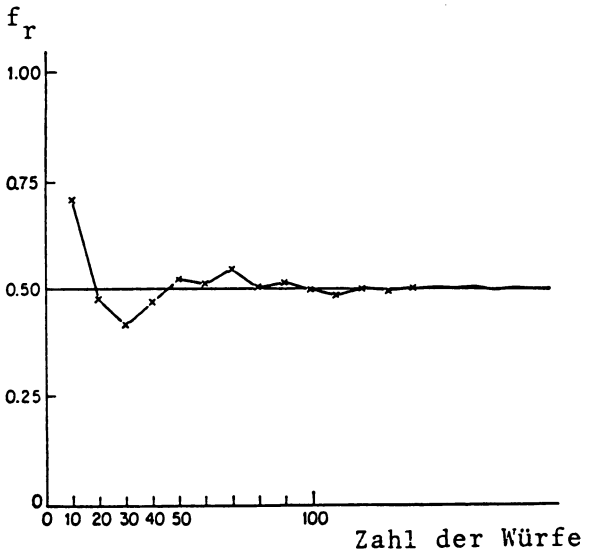
chung (6-14)). Das Konzept der Unabhängigkeit zweier Ereignisse ist die Basis für die Definition der Unabhängigkeit zweier Zufallsvariablen. Man betrachtet zwei diskrete Zufallsvariablen Y und X mit den Ausprägungen y_i ($i = 1,2, \ldots, k$) und x_j ($j = 1,2, \ldots, l$) als stochastisch unabhängig voneinander, wenn gilt: $p(y_i + x_j) = p(y_i)\ p(x_j)$ für **alle** $i = 1,2, \ldots, k$ und **alle** $j = 1,2, \ldots, l$. Im Falle der statistischen Unabhängigkeit stimmen alle bedingten Verteilungen mit den entsprechenden Randverteilungen überein (vergl. hierzu Teil I, Abschnitt 4.2.2). Wenn man die Unabhängigkeit zweier **kontinuierlicher** Zufallsvariablen in Begriffen der Wahrscheinlichkeitstheorie definieren will, muß man auf die Dichtefunktionen zurückgreifen: Eine Zufallsvariable X mit der Dichte g(x) und eine Zufallsvariable Y mit der Dichte h(x) sind unabhängig voneinander, wenn für alle Kombinationen (x^*, y^*) gilt: $f(x^*, y^*) = g(x^*)\ h(y^*)$, wobei $f(x^*, y^*)$ die sog. **gemeinsame Dichte** für das Ereignis $(X = x^*; Y = y^*)$ darstellt.

$$(6\text{-}40) \qquad F(x,y) = P(X \leq x^* \text{ und } Y \leq y^*) = \int_{-\infty}^{x^*} \int_{-\infty}^{y^*} f(x,y)\,dx\,dy$$

Auch mehrdimensionale Wahrscheinlichkeitsverteilungen lassen sich durch bestimmte Kennwerte charakterisieren. Zu den Kennwerten für die jeweiligen Randverteilungen treten jetzt noch die Kennwerte für den Grad der Abhängigkeit der beiden Variablen. Dazu haben wir einige empirische Kenngrößen (Kovarianzen bzw. Korrelationskoeffizienten im weitesten Sinne) für Variablen unterschiedlichen Meßniveaus in Teil I, Kap. 4 besprochen. Weitere werden wir in Kap. 10 noch kennenlernen.

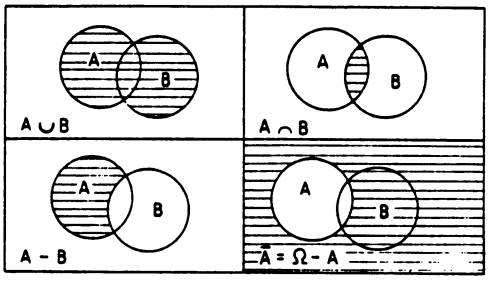
So wie wir Erwartungswerte und Varianzen für eindimensionale Wahrscheinlichkeitsverteilungen analog zu den entsprechenden empirischen Kenngrößen definiert haben, lassen sich auch die theoretischen Kenngrößen für den Grad des Zusammenhangs zweier Zufallsvariablen aus den entsprechenden empirischen Kenngrößen herleiten. Wir wollen dies hier jedoch nicht weiter ausführen.

Abb. 6.1: Relative Häufigkeiten von "Wappen" bei unterschiedlich langen Sequenzen von Münzwürfen



Quelle: Blalock 1960, S. 99

Abb. 6.2: Mengendiagramme



Quelle: Hartung 1986, S. 92

Abb. 6.3: Formale Schulbildung und Mobilitätsgrad der Reichstagsabgeordneten von 1912

	Formale Schulbildung niedrig 1	mittel 2	hoch 3	
Mobilität:				
niedrg 1	35	26	103	164
mittel 2	38	15	113	166
hoch 3	15	11	61	87
	88	52	277	417

Abb. 6.4: Wahrscheinlichkeitsfunktion für eine diskrete Zufallsvariable

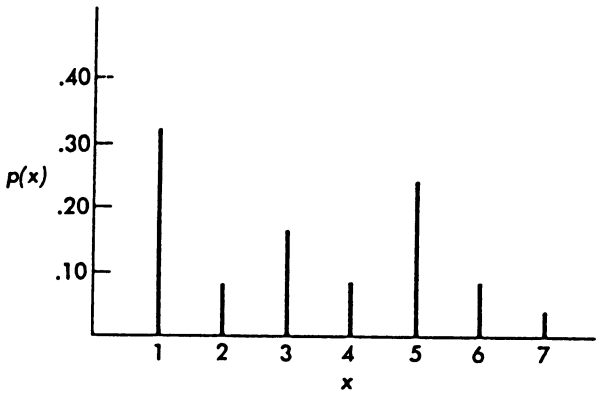


Abb. 6.5: Veranschaulichung des Übergangs von der relativen Häufigkeitsdichte zur Wahrscheinlichkeitsdichte

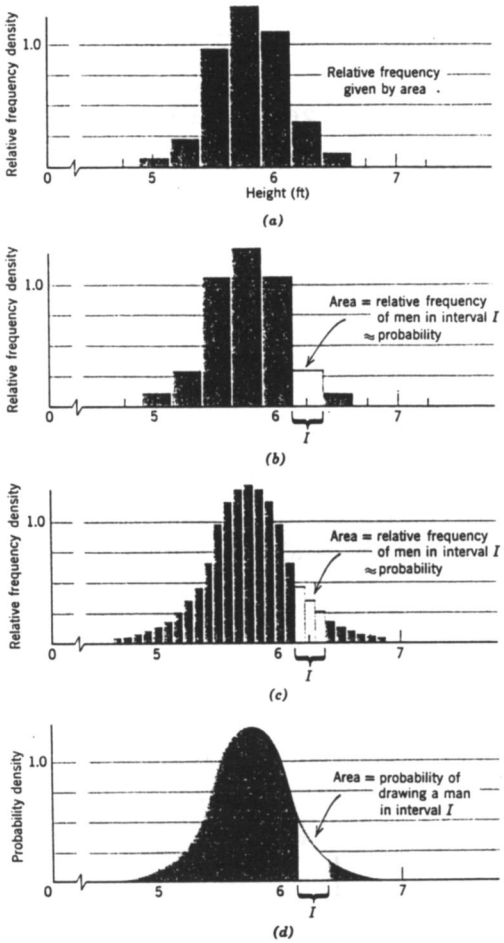


FIGURE 4-7 How relative frequency density may be approximated by a probability density as sample size increases, and cell size decreases. (a) Small *n*, as in Fig. 4-6*b*. (*b*) Large enough *n* to stabilize relative frequencies. (c) Even larger *n*, to permit finer cells while keeping relative frequencies stable. (*d*) For very large *n*, this becomes (approximately) a smooth probability density curve.

Quelle: Wonnacott/Wonnacott 1972, S. 73

Abb. 6.6: Verteilungsfunktion einer diskreten Zufallsvariablen

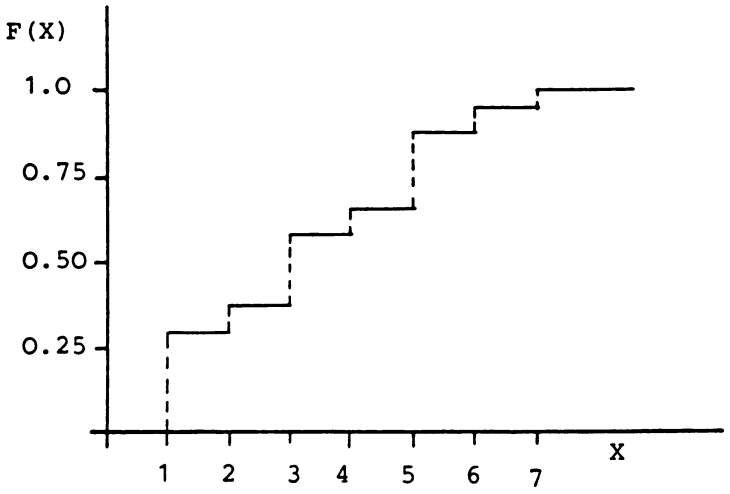
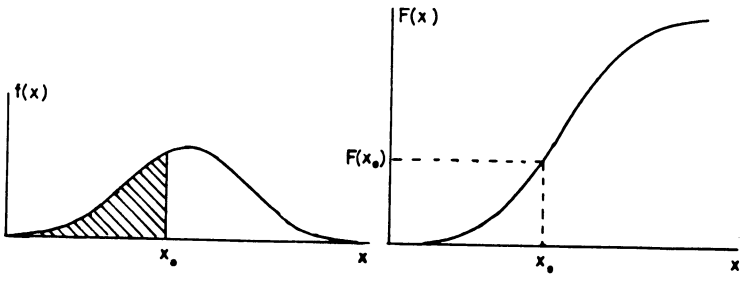
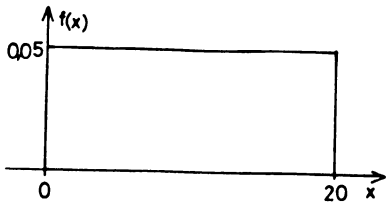


Abb. 6.7: Dichtefunktion f(x) und Verteilungsfunktion F(x) einer kontinuierlichen Zufallsvariablen



Quelle: Schlittgen 1987, S. 97

Abb. 6.8: Dichtefunktion von Wartezeiten



Quelle: Schlittgen 1987, S. 97

Abb. 6.9: Sterbewahrscheinlichkeiten für 9 Lebensdekaden

i	1	2	3	4	5	6	7	8	9
x_i	0	1	2	3	4	5	6	7	8
p_i	0.017	0.021	0.029	0.090	0.217	0.328	0.239	0.058	0.001

Quelle: Schlittgen 1987, S. 124

KAPITEL 7

Stichprobenfunktionen und ihre Verteilungen

7.1. Zum Konzept der Stichprobenfunktion und der Stichprobenverteilung

Stellen wir uns folgende Forschungssituation vor: Ein Sozialhistoriker möchte die materiellen Lebensbedingungen der Berliner Arbeiterhaushalte von 1900 untersuchen. Nehmen wir an, aus vorliegenden Zensusdaten ließe sich u. a. das Monatseinkommen sämtlicher Berliner Arbeiterhaushalte dieser Zeit rekonstruieren. Um unnötigen Geld- und Zeitaufwand zu sparen, entschließt sich unser Forscher, keine Totalerhebung vorzunehmen, sondern nach dem Zufallsprinzip (siehe Kap. 9) 2000 Arbeiterhaushalte auszuwählen und deren Einkommen zu notieren. Wir wollen annehmen, daß ihm dieses Vorhaben so gelingt, daß jeder Berliner Arbeiterhaushalt von 1900 die gleiche Chance hat, in die »Stichprobe« mit n = 2000 Fällen zu gelangen, und daß jede »Ziehung« eines Haushalts unabhängig ist von der Ziehung eines anderen Haushalts. Nehmen wir weiterhin an, in der »Grundgesamtheit«, die alle Berliner Arbeiterhaushalte bilden, sei die in Abb. 7.1 dargestellte (fiktive) Einkommensverteilung f(x) gegeben. Jede einzelne Haushaltsziehung kann nun als Zufallsexperiment angesehen werden mit der Zufallsvariable X_i = Monatseinkommen der Arbeiterhaushalte i, i = 1, 2, . . ., n. Wir haben es also mit n Zufallsvariablen zu tun, die voneinander unabhängig sind. Wir bezeichnen sie als **Stichprobenvariablen.** Jede dieser Zufallsvariablen hat die gleiche Wahrscheinlichkeitsverteilung, die sich nach dem Gleichmöglichkeitsmodell aus den relativen Häufigkeiten der Grundgesamtheit ergibt[1]. Auch wenn wir diese nicht kennen, legen sie natürlich die Wahrscheinlichkeiten dafür fest, daß bei der Stichprobenvariablen X_i ein bestimmter Wert realisiert wird. Wenn z. B. 60 % der Arbeiterhaushalte über ein Monatseinkommen von 900 bis 1100 Reichsmark verfügen, gibt es bei jeder Zufallsziehung eines Haus-

[1] Streng genommen setzen wir dabei voraus, daß ein »gezogener« Haushalt vor der Ziehung des nächsten Haushalts wieder in die Grundgesamtheit »zurückgelegt« wird. In der Praxis wird normalerweise nicht so verfahren. Die Ungenauigkeiten, die durch die Veränderungen des Stichprobenraumes entstehen, sind aber vernachlässigbar, wenn die Grundgesamtheit im Verhältnis zur Stichprobe »sehr groß« ist. Siehe hierzu auch unten S. [42f] sowie S. [49], Fn. 5.

halts aus dieser Grundgesamtheit eine Wahrscheinlichkeit von p = 0,6, daß er monatlich einen Betrag zwischen 900 und 1100 RM einnimmt (siehe Gleichung (6-1)).

Wir haben oben ausgeführt, daß für kontinuierliche Zufallsvariablen keine »Wahrscheinlichkeitsfunktion« definiert ist (sondern eine »Wahrscheinlichkeitsdichtefunktion«), daß aber sowohl stetige als auch diskrete Zufallsvariablen durch ihre (kumulierten) Verteilungsfunktionen F(x) charakterisiert sind. Wir können deshalb allgemein definieren: Eine einfache Zufalls-**Stichprobe** aus einer Grundgesamtheit, in der die Variable X die Verteilungsfunktion F(x) hat, besteht aus n Zufallsvariablen $X_1, \ldots, X_n$, den sog. Stichprobenvariablen, die voneinander unabhängig sind und alle dieselbe Verteilungsfunktion F(x) haben. Man spricht auch von einer »Stichprobe aus der Verteilung F(x)«. Die Stichprobenziehung bedeutet also die wiederholte Durchführung eines gleichbleibenden Zufallsexperiments.

In der Regel möchte der Forscher von einer Verteilung, die er in der Stichprobe beobachtet, auf die (meistens unbekannte) Verteilung der Grundgesamtheit schließen (»induktive« Schlußweise). Wir bleiben aber zunächst bei der deduktiven Betrachtungsweise. Später (in Kap. 8) wird das induktive Schließen als Anwendung der deduktiven Logik »in umgekehrter Richtung« plausibel werden.

Meistens ist der Forscher nicht an den detaillierten Häufigkeitsverteilungen der Grundgesamtheit interessiert, sondern nur an bestimmten summarischen Kennwerten, z. B. dem arithmetischen Mittel oder einem Korrelationskoeffizienten. Wir müssen also fragen, ob es möglich ist, auch für solche (zu schätzende) Kennwerte Wahrscheinlichkeitsverteilungen (bzw. Verteilungsfunktionen) anzugeben.

Daß wir aus den realisierten Stichproben **empirische** Kennwerte berechnen können, ist selbstverständlich. Zu klären ist, in welcher Weise und auf welcher Grundlage wir Stichprobenkennwerte als Schätzgrößen für die entsprechenden Kennwerte (»Parameter«) der empirischen oder hypothetischen Population benutzen können.

Ein erster Schritt zur Lösung dieses Problems ist das Konzept der **Stichprobenfunktion.** Auch wenn unsere praktischen Möglichkeiten stark begrenzt sind, so können wir uns doch wenigstens vorstellen, nicht nur eine, sondern mehrere Stichproben mit jeweils n Fällen zu ziehen. Dann würden mit den jeweils realisierten Werten $x_1, \ldots, x_n$ der Stichprobenvariablen $X_1, \ldots, X_n$ auch die Werte der jeweiligen empirischen Kennwerte von Stichprobe zu Stichprobe variieren. Zu den einzelnen Stichproben würden wir z. B. unterschiedliche, in bestimmten Grenzen streuende, arithmetische Mittel erhalten. Man kann also die Kennwerte selbst als Zufallsvariablen auffassen, die sich aus Funktionen der Stichprobenvariablen ergeben. So ist das arithmetische Mittel nichts anderes als die gewichtete

Summe der einzelnen Werte: $\bar{x} = 1/n(x_1 + x_2 + ... + x_n)$. Der Schritt, die Kennwerte als Stichprobenfunktion $g(X_1, \ldots, X_n)$ und damit als Zufallsvariable aufzufassen, ist grundlegend für alle weiteren Überlegungen, die wir zur Inferenzstatistik noch anstellen werden.

Wenn Stichprobenfunktionen Zufallsvariablen sein sollen, dann muß es für sie auch Wahrscheinlichkeitsverteilungen geben. Man bezeichnet die Wahrscheinlichkeitsverteilung einer Stichprobenfunktion als **Stichprobenverteilung** (englisch: »sampling distribution«), gelegentlich spricht man auch von »Stichprobenkennwertverteilungen«. Im Prinzip können die Verteilungen verschiedener Stichprobenfunktionen aus den Verteilungen der zugrunde liegenden Stichprobenvariablen abgeleitet werden. Gelegentlich ist das nur näherungsweise möglich. Für kleine Stichproben werden die Stichprobenverteilungen oft auch tatsächlich durch Experimente bzw. Simulationen am Computer geschätzt (sog. Monte-Carlo-Studien). Wir wollen jetzt anhand eines einfachen Beispiels verdeutlichen, wie man (theoretische) Stichprobenverteilungen aus den Verteilungen der Stichprobenvariablen ableiten kann.

7.2 Binomial- und Multinomialverteilung

Zu diesem Zweck betrachten wir zunächst Stichprobenvariablen, bei denen nur interessiert, ob bei Ablauf eines Zufallsexperiments ein bestimmtes Ereignis A (ein bestimmter Wert x) realisiert wird oder nicht. Ein Beispiel hierfür ist das Werfen einer Münze, wobei nur wichtig ist, ob »Wappen« (oder »Zahl«) oben liegt. Ein anderes Beispiel ist die Zufallsauswahl aus der Grundgesamtheit der Reichstagsabgeordneten, wenn bei der Variablen »X: = Fraktionszugehörigkeit der Abgeordneten« nur interessiert, ob sie dem Zentrum (oder einer anderen Partei, für die man sich gerade interessiert) angehören oder nicht. Man spricht in diesem Falle von »binär« kodierten Variablen: Dem interessierenden Ereignis (hier: Zentrumsabgeordneter sein) ordnet man in der Regel die Zahl 1, dem anderen (komplementären) Ereignis die Zahl 0 zu. Das erste Ereignis wird im folgenden mit Z, das zweite mit $\bar{Z}$ (Nicht-Z) bezeichnet.

Wir wollen für unser Beispiel annehmen, daß die Zufallsexperimente (Stichprobenziehungen) unabhängig voneinander sind und daß die (»Erfolgs«-)Wahrscheinlichkeit für das Ereignis Z, $P(Z) = p$, konstant ist (d. h., die jeweils ausgewählten Abgeordneten werden vor der nächsten Ziehung wieder in die Grundgesamt eingeordnet). Eine solche Serie von unabhängigen Zufallsvorgängen mit binären Variablen bezeichnet man als **Bernoulli-Prozeß**. Der summarische Kennwert, die Stichprobenfunktion, die hierbei vor allem interessiert, ist die Summe der Realisationen von Z bei n Versuchen - im Beispiel: Wie groß ist die Zahl der Zentrumsab-

geordneten in einer Zufallsstichprobe von n = 100 Fällen? Die Stichprobenvariable »X:= Fraktionszugehörigkeit der Reichstagsabgeordneten« hat zwei Ausprägungen: 1:= Zentrumszugehörigkeit, 0:= keine Zugehörigkeit zum Zentrum. Die Stichprobenfunktion als Zufallsvariable »S:= Anzahl der Zentrumsabgeordneten bei 100 Zufallsziehungen« hat dagegen 101 Realisationsmöglichkeiten: 0, 1, 2, ..., 100. Nehmen wir an, wir wüßten bereits, daß in der Grundgesamtheit der Reichstagsabgeordneten von 1871 bis 1914 der Anteil der Zentrumspolitiker 20 % beträgt. Dann dürfte es »sehr unwahrscheinlich« sein, in einer Zufallsstichprobe mit n = 100 Ziehungen nur zwei oder drei Zentrumsabgeordnete vorzufinden. Ebenso unwahrscheinlich wäre es, 97 oder 98 Zentrumsabgeordnete auszuwählen. Aber wir müssen nicht bei so ungenauen Aussagen stehenbleiben. Wenn wir die Stichprobenziehung nach dem Modell des Bernoulli-Prozesses gestalten und die relative Häufigkeit der Zentrumsabgeordneten in der Grundgesamtheit kennen, läßt sich die Wahrscheinlichkeitsverteilung (Stichprobenverteilung) der Zufallsvariablen S genau bestimmen:

Um den Rechenvorgang zu erleichtern, demonstrieren wir die Ableitung der Stichprobenverteilung für einen kleineren Stichprobenumfang n = 4. Folglich gibt es für die Stichprobenfunktion S nur noch 5 mögliche Werte (Ausprägungen): 0, 1, ..., 5. Die relative Häufigkeit der Zentrumsabgeordneten in der Grundgesamtheit nehmen wir weiterhin mit $f_r = 0{,}20$ an. Nach dem klassischen Gleichmöglichkeitsmodell (s. Kap. 1.6) ergeben sich daraus die Wahrscheinlichkeiten für die Stichprobenvariablen X_i, $i = 1, \ldots, 4$ jeweils mit $p(X=1)=0{,}20$ und $p(X=0)=(1-0{,}20)=0{,}80$. Die Wahrscheinlichkeitsverteilung für die Stichprobenfunktion $S = f(X_1, X_2, X_3, X_4) = X_1 + X_2 + X_3 + X_4$ läßt sich wie in Abb. 7.2 dargestellt ableiten (Z = Auswahl eines Zentrumsabgeordneten, $\overline{Z}$ = Auswahl eines Abgeordneten, der nicht dem Zentrum angehört).

Die Wahrscheinlichkeiten ergeben sich aus der Voraussetzung $p(X_i = 1) = 0{,}20$ und, daraus folgend, $p(X_i = 0) = 0{,}80$ sowie der Anwendung des Additions- und des Multiplikationstheorems (siehe Kap. 6). Betrachten wir als Beispiel die Wahrscheinlichkeitsangaben im 2. Zeilenblock. Jede der dort genannten Ereignisfolgen enthält in unterschiedlichen Anordnungen dreimal den Wert $X_i = 0$ und einmal den Wert $X_i = 1$. Folglich ergibt sich jedesmal nach dem Multiplikationssatz (Verknüpfung unabhängiger Ereignisse) für sie eine Wahrscheinlichkeit von $p = 0{,}2 \cdot (1-0{,}2)^3$. Alle 4 Einzelergebnisse realisieren für die Stichprobenfunktion S (Summenvariable) den Wert S = 1. Nach dem Additionssatz (Verknüpfung disjunkter Ereignisse) erhalten wir also für die Summe S = 1 die Wahrscheinlichkeit

$$p(S=1) = 4 \cdot 0{,}2(1-0{,}2)^3.$$

Wir wollen dieses Beispiel jetzt auf beliebige Bernoulliprozesse mit n Versuchen (wiederholten »Ziehungen«) verallgemeinern. Die Zufallsvariable S (als Stichprobenfunktion) beinhalte weiterhin als mögliche Werte die Häufigkeit, mit der das Ereignis A bei einer Serie von n Versuchen auftreten kann. Die Wahrscheinlichkeiten für die verschiedenen Häufigkeiten $s \leqq n$ sollen ermittelt werden unter der Voraussetzung, daß bei jeder einzelnen Ziehung $P(A) = p$ und $P(\bar{A}) = (1-p)$ gilt. Eine Versuchsserie produziert eine bestimmte Folge der binär kodierten Ereignisse, in denen das Ereignis A s-mal und das Ereignis Nicht-A (n-s)mal auftritt, beispielsweise die Folge $(\bar{A}\bar{A}A\bar{A}A \ldots A\bar{A})$. Wegen der (vorausgesetzten) Unabhängigkeit der einzelnen Versuche gilt: $P(\bar{A}\bar{A}A\bar{A}A \ldots A\bar{A}) = P(\bar{A})\,P(\bar{A})\,P(A)\,P(\bar{A})\,P(A) \cdot \ldots \cdot P(A)\,P(\bar{A}) = p^s(1-p)^{n-s}$ (Anwendung des Multiplikationstheorems). Eine andere Abfolge, eine andere »Kombination« von s Ereignissen A und (n-s) Ereignissen Nicht-A ($\bar{A}$) hat die gleiche Wahrscheinlichkeit. Letztlich interessiert uns aber nur die **Häufigkeit** s, mit der das »günstige« Ereignis A in einer Serie von n Versuchen auftritt, nicht die jeweilige Abfolge der A- und $\bar{A}$-Ereignisse. Um die Wahrscheinlichkeit $P(S = s)$ für eine bestimmte Häufigkeit von A ermitteln zu können, müssen wir also herausfinden, wieviele (gleichwahrscheinliche) Kombinationen von s A- und (n-s)$\bar{A}$-Ereignissen überhaupt möglich sind. Laut Gleichung (6-18") sind es

$$\frac{n_j}{s!\,(n-s)!} = \binom{n}{s}$$

Möglichkeiten. Deshalb erhielten wir in unserem obigen Beispiel mit den Zentrumsabgeordneten $\binom{4}{1} = 1 \cdot 2 \cdot 3 \cdot 4 / 1 \cdot 1 \cdot 2 \cdot 3 = 4$ Ergebnisse mit jeweils $S = 1$. (Die Größe $\binom{n}{s}$ bezeichnet man in diesem Kontext als **Binomialkoeffizient**).

Die Wahrscheinlichkeit, bei einem aus n Versuchen bestehenden Bernoulli-Prozeß genau s-mal das Ereignis A zu erhalten, ist also

$$\binom{n}{s} p^s(1-p)^{n-s}. \tag{7-1}$$

Demgemäß heißt allgemein eine Zufallsvariable X binomialverteilt mit den Parametern n und p, kurz: $X \sim B(n,p)$, wenn die Wahrscheinlichkeitsfunktion von X gegeben ist durch

$$(7\text{-}2) \qquad f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , \; x = 0,1,\ldots,n.$$

Wie vorher »s«, so gibt hier »x« die Häufigkeit des interessierenden Ereignisses an. In der Praxis muß man sich diese Verteilung nicht stets selbst ausrechnen, da sie in tabellierter Form in vielen Lehrbüchern (z. B. Schlittgen 1987) vorliegt.

Die Gestalt der Binomialverteilung hängt von der Wahrscheinlichkeit ab, mit der das interessierende Ereignis A bei einem einzigen Versuch auftreten kann. Symmetrisch ist die Verteilung nur bei $P(A) = 0{,}5$. Die Abbildung 7.3 zeigt weitere Varianten in Abhängigkeit von den Parametern n und p.

Die Binomialverteilung nähert sich der symmetrischen Form um so stärker an, je näher p bei dem Wert 0,5 liegt und je größer n ist.

Da wir inzwischen gewohnt sind, Verteilungen durch Lage- und Streuungsparameter zu charakterisieren, wollen wir nun auch den Erwartungswert E(X) und die Varianz V(X) einer binomialverteilten Zufallsvariablen X (Stichprobenfunktion) bestimmen. Dabei setzen wir voraus, daß die entsprechende Stichprobenvariable X_i ($i = 1,2, \ldots, n$) binär im obigen Sinne kodiert ist: $X_i = 1$, falls im i-ten Versuch das interessierende Ereignis A eintritt; $X_i = 0$, falls im i-ten Versuch das Komplementärereignis $\overline{A}$ eintritt. Folglich gilt

$$P(X_i = x_i) = P(X_i = 1) = p$$

und

$$P(X_i = x_i) = P(X_i = 0) = (1 - p).$$

Außerdem definieren wir die Stichprobenfunktion X (in unserem obigen Beispiel mit »S« bezeichnet) als eine Zufallsvariable $X = \Sigma X_i$. Aus diesen Festlegungen resultiert

a) für die Stichprobenvariable X_i:

(7-3) $$E(X_i) = \sum_{v=0}^{1} x_{iv}p_v = 1\cdot p + 0(p-1) = p \quad , \quad v = \text{Index der Ausprägungen}$$

(gilt für beliebige i, i = 1,2,...,n)

$$V(X_i) = \sum_{v=0}^{1} (x_{iv} - p)^2 p_v = (1-p)^2 p + (0-p)^2(1-p)$$
$$= p(1-p)$$

b) für die Stichprobenfunktion X:

(7-4) $$E(X) = E(\sum_{i=1}^{n} X_i) = \sum E(X_i) = n\cdot p \; ; \text{ siehe (6-31)}$$

$$V(X) = V(\sum X_i) = \sum V(X_i) = np(1-p),$$

da die X_i unabh. sind und wegen (6-39)

Folglich können wir bei n = 100 zufälligen Ziehungen aus der **Grundgesamtheit** der Reichstagsabgeordneten erwarten, 100·0,20 = 20 Zentrumsabgeordnete zu erhalten, wenn p(Z) = 0,2[2]

Allgemein gilt: Falls X und Y unabhängige binomialverteilte Zufallsvariablen mit gleichem Parameter p darstellen, X ~ B(n,p) und Y ~ B(m,p), ist ihre Summe Z ebenfalls binomialverteilt: Z = (X + Y) ~ B(n + m, p).

Für viele Forschungsfragen ist eine Beschränkung auf binär kodierte Zufallsvariablen natürlich nicht einzuhalten. Oft interessieren nicht nur zwei Ereignistypen (die wechselseitig komplementär zueinander sind), sondern k > 2 Ereignistypen. Wenn für jeden Typ ermittelt werden soll, wie oft er bei n Versuchen realisiert wird, haben wir es mit mehreren Stichprobenfunktionen X_1, X_2, ..., X_k zu tun. Jede dieser Zufallsvariablen X_i, i = 1,2, . . ., k, beinhalte die Anzahl x_i der Elemente (Ereignisse) des

[2] Der Erwartungswert $E(X_i) = p$ für die binäre Stichprobenvariable ist zwar keine anschauliche Größe (man wird ja nicht 0,2 Zentrumsabgeordnete bei einer Ziehung »erwarten«), man braucht diese formale Bestimmug aber um den (dann auch anschaulichen) Erwartungswert für die Stichprobenfunktion ableiten zu können.

iten Typs in der Serie von n Versuchen; p_i sei der Anteil des iten Typs in der Grundgesamtheit. Es gilt dann (bei sonst gleichen Voraussetzungen wie im Bernoulliprozeß) für diese sog. **Multinomialverteilung**:

$$(7\text{-}5)\quad P(X_1=x_1,\ X_2=x_2,\ \ldots,X_k=x_k) = \frac{n!}{x_1!x_2!\ldots x_k!}\, p_1^{x(1)}\ldots p_k^{x(k)}$$

$$\text{mit } i = 1,\ldots,k;\quad \sum x_i = n\ ;\ \sum p_i = 1$$

Wir sehen also, daß die Binomialverteilung ein spezieller Fall (mit $k=2$) der Multinomialverteilung ist. Wir haben hier den speziellen Fall abgeleitet, weil er rechentechnisch einfacher zu handhaben ist und dennoch das Ableitungsprinzip hinreichend deutlich werden läßt.

Die hier erläuterte Ableitung der Binomial- und Multinomialverteilung setzt voraus, daß sich die Wahrscheinlichkeiten für das Auftreten der jeweiligen Ereignisse nicht nach Eintritt eines Ereignisses verändern. Wenn wir aber aus einer **kleinen** Grundgesamtheit eine Stichprobe ziehen, ohne den »gezogenen« Fall wieder in die Grundgesamtheit »zurückzulegen«, ergeben sich bei der nächsten Ziehung mehr oder weniger veränderte Wahrscheinlichkeiten für die jeweils möglichen Ergebnisse. Nehmen wir z. B. eine Grundgesamtheit von $N=100$ Parlamentarieren an, in der sich 20 Abgeordnete des Zentrums befinden. Es soll eine Stichprobe von $n=10$ Abgeordneten gezogen werden. Vor der 1. Ziehung gibt es eine Wahrscheinlichkeit von $p_1 = 0{,}2$, daß ein Zentrumsabgeordneter ausgewählt wird. Falls sich diese Möglichkeit schon beim ersten Zug realisiert, gibt es vor »Ziehung« des zweiten Abgeordneten eine Wahrscheinlichkeit von $p_2 = 19/99 = 0{,}192$, daß erneut ein Zentrumsabgeordneter ausgewählt wird. Wenn man bei binären oder multinomialen Zufallsvariablen nicht das Modell einer »Ziehung mit Zurücklegen« bzw. einer Ziehung aus einer unendlichen Grundgesamtheit unterstellen kann und statt dessen von dem Modell einer »Ziehung ohne Zurücklegen« ausgehen muß, führt dies zum **hypergeometrischen** bzw. zum multidimensionalen hypergeometrischen Verteilungsmodell anstelle der Binomial- bzw. der Multinomialverteilung. Knappe Erläuterungen hierzu geben z. B. Kriz (1973, S. 92 ff.); Hartung et al. (1986, S. 207 f.)

Es sollte erkennbar geworden sein, daß die vielfältigen Forschungsfragen, die auftreten können, jeweils eine spezifische Definition von Zufallsvariablen implizieren. Für die allermeisten dieser Zufallsvariablen haben die Statistiker bereits geeignete theoretische Verteilungen ausgearbeitet,

die wir hier natürlich nicht alle vorstellen, geschweige denn ableiten können. Eine, nämlich die Binomialverteilung, sollte hier aber Schritt für Schritt abgeleitet werden, um zu zeigen, wie solche theoretischen Verteilungen für Stichprobenfunktionen (Kennwerte) im Prinzip konstruiert werden können. Dadurch sollte auch die konzeptuelle Unterscheidung von empirischen Häufigkeitsverteilungen und theoretischen Wahrscheinlichkeitsverteilungen noch einmal verdeutlicht werden.

Wir benötigen Stichprobenverteilungen vor allem, um statistische Kennwerte für Populationen aus Stichprobendaten schätzen oder entsprechende Hypothesen testen zu können (siehe Kap. 8). Eine weitere theoretische Verteilung soll ebenfalls etwas ausführlicher besprochen werden, weil sie eine besonders wichtige Rolle spielt: die Normalverteilung.

7.3 Die Normalverteilung

Kehren wir noch einmal zu unserem Ausgangsbeispiel, der Population Berliner Arbeiterhaushalte und der Zufallsvariable X = Monatliches Haushaltseinkommen zurück. Stellen wir uns vor, wir würden aus dieser relativ großen Grundgesamtheit von etwa N = 500000 Fällen k relativ kleine Stichproben von $n_i = 1000$ Fällen ($i = 1, \ldots, k$) nach dem Zufallsprinzip auswählen. Für jede Stichprobe ließe sich ein arithmetisches Mittel $\bar{x}_i$ berechnen. Man kann vermuten,

- daß die arithmetischen Mittel $\bar{x}_i$ der Stichproben um das arithmetische Mittel μ der Grundgesamtheit einigermaßen symmetrisch streuen,
- daß, bei sehr vielen Stichprobenziehungen, die $\bar{x}_i$ mit geringem Abstand zu μ häufiger vorkommen als die $\bar{x}_i$ mit großem Abstand zu μ,
- daß sich das arithmetische Mittel der Stichprobenmittel um so stärker dem Parameter μ annähert, je größer n bzw. je größer die Zahl k der Stichproben.

Mit anderen Worten: man kann mit einer glockenförmigen Dichtefunktion für die Stichprobenfunktion »arithmetisches Mittel« rechnen; sie ist um so weiter auseinandergezogen, je geringer der Umfang der Stichproben ist. Mathematiker bzw. Statistiker sind natürlich in der Lage, diese Überlegungen zu präzisieren und formal zu begründen. So haben sie herausgefunden, daß das arithmetische Mittel unter bestimmten Bedingungen (die wir noch kennenlernen werden) »normalverteilt« ist. Auch einige andere Stichprobenfunktionen sind unter bestimmten Bedingungen normalverteilt. Die (glockenförmige) Dichtefunktion einer normalverteilten Zufallsvariablen X mit dem arithmetischen Mittel μ und der Standardabweichung σ (abgekürzt: $X \sim N(\mu,\sigma)$ ist allgemein gegeben durch

$$(7\text{-}6) \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Die Größen $\pi \approx 3{,}14$ (Verhältnis des Kreisumfangs zum Durchmesser) und $e \approx 2{,}718$ (Eulersche Zahl, Basis der natürlichen Logarithmen) sind gegebene Konstanten. Der Lageparameter μ (Erwartungswert) und der Streuungsparameter σ (die Standardabweichung) halten die Dichtefunktion variabel. Abb. 7.4 zeigt einige Beispiele von Normalverteilungen, bei denen Mittelwert und Standardabweichung variieren.
Es ist nicht möglich, die Herleitung der Gleichung (7-6) an dieser Stelle zu erklären. Wir wollen aber einige ihrer formalen Eigenschaften hervorheben, die für die Praxis wichtig sind:

Für $x = \mu$ ist der Exponent null. Wegen des negativen Vorzeichens ist er an allen anderen Stellen kleiner als Null. Also hat die Dichte an der Stelle $x = \mu$ ihr Maximum (Scheitelpunkt). Da man theoretisch von einem unendlich großen Wertebereich ausgeht, nähern sich die Kurvenenden asymptotisch der X-Achse. Wegen des Quadrierens führen positive und negative Differenzen $(x - \mu)$ zum gleichen Exponenten, also zu gleichen Dichtewerten. Die Kurve ist somit symmetrisch um μ. Obwohl die einzelnen Normalverteilungskurven je nach Parameterkonstellation[2a] von μ und σ unterschiedlich stark gestreckt oder gestaucht sind, gilt für alle Dichtefunktionen der Normalverteilung, daß ihre Wendepunkte bei denjenigen X-Werten liegen, die um 1 Standardabweichung (positiv und negativ) vom Mittelwert entfernt sind. Die Fläche, die von den beiden Ordinaten an den Stellen $-\sigma$ und $+\sigma$ eingeschlossen wird, hat bei jeder Normalverteilung (unabhängig von der jeweiligen Parameterkonstellation) einen Anteil von 68,26 % an der Gesamtfläche (siehe Abb. 7.5). Die Flächenanteile zwischen zwei Ordinaten an beliebig festgehaltenen Stellen $x_1 = a \cdot \sigma$ und $x_2 = b \cdot \sigma$ (a, b sind beliebig wählbare Konstanten ungleich Null) sind für alle Normalverteilungen gleich, unabhängig davon, wie groß die jeweilige Standardabweichung σ ist. Das werden wir unten noch weiter ausführen.

Bevor wir uns näher mit den formalen Eigenschaften der Normalverteilung beschäftigen, wollen wir bemerken, daß keineswegs nur das arithmetische Mittel wiederholter Stichprobenauswahlen normalverteilt ist.

[2a] In der Literatur wird sowohl die Varianz als auch ihr Wurzelausdruck, die Standardabweichung, als Streuungsparameter eingesetzt. Folglich findet man $X \sim N(\mu, \sigma^2)$ oder $X \sim N(\mu, \sigma)$.

Schlittgen (1987, S.218) führt aus, daß Zufallsvariablen »oft« normalverteilt sind, wenn auf sie folgender Sachverhalt zutrifft:

- »Die Variable beschreibt eine natürliche Variation wie Körpergröße, Gewicht, Länge der Blätter eines Baumes.
- Die Variable beschreibt das Ergebnis einer Messung einer physikalischen Größe wie z. B. der Länge eines Raumes oder der Meßgenauigkeit einer Waage.
- Die Variable entsteht durch die Summe unterschiedlicher Zufallseinflüsse. Dies gilt etwa für den Intelligenzquotienten, der sich aus den Punkten vieler einzelner Fragen ergibt, . . .«

Ursprünglich wurde die Normalverteilung entwickelt, um das Auftreten zufälliger Meß- und Beobachtungsfehler in Situationen zu beschreiben, in denen man einen »wahren« Wert unterstellen konnte. Eine solche Situation war z. B. bei der Beobachtung von Planetenbahnen gegeben. Die Normalverteilung wird gelegentlich nach dem Mathematiker C. F. Gauß »Gaußsche Glockenkurve« genannt (obwohl sie von DeMoivre entdeckt wurde).

Mathematisch gesehen, steht die Normalverteilung der Binomialverteilung viel näher, als es auf den ersten Blick bei einem Vergleich der Funktionen (7-2) und (7-6) erscheinen mag. Je größer der Stichprobenumfang n wird, umso stärker nähert sich die Binomialverteilung der Normalverteilung an, und zwar um so rascher, je näher der Anteilswert p (und damit auch 1 - p) bei 0,5 liegt. Wir hatten ja schon weiter oben angemerkt, daß die Binomialverteilung um so symmetrischer wird, je größer n ist. Als Faustregel kann davon ausgegangen werden, daß die Approximation der Binomialverteilung durch die Normalverteilung hinreichend gut ist, wenn $n \cdot p \geqq 10$ und $n(1 - p) \geqq 10$ ist (Schlittgen 1987, S.231)[3]. Dies zu wissen ist gelegentlich von Nutzen, weil für große n die Binomialverteilung nicht mehr tabelliert ist.

Die herausragende Bedeutung der Normalverteilung in der Statistik hat im wesentlichen zwei Gründe (ebd., S.218):

(a) Sie kann bei einer Vielzahl von statistischen Maßzahlen (wie z. B. dem arithmetischen Mittel) als Verteilungsmodell unterstellt werden, wenn nur die Stichproben genügend groß sind (siehe unten).
(b) Sie weist einige formale Eigenschaften auf, die sie außerordentlich »praktikabel« machen.

Neben den eingangs schon erwähnten sind für unsere Zwecke vor allem folgende Eigenschaften wichtig:

---

[3] Einige andere Autoren nennen als Faustregel: $np(1-p) \geqq 9$.

(1) Eine Variable $Y = a + bX$, die aus einer Lineartransformation einer normalverteilten Variablen, $X \sim N(\mu,\sigma)$, hervorgegangen ist, ist ebenfalls normalverteilt: $Y \sim N(a + b\mu, |b|\sigma)$.
Ihr Erwartungswert und ihre Varianz bzw. Standardabweichung lassen sich durch Anwendung der Rechenregeln für Erwartungswerte ermitteln:

$$(7\text{-}7) \qquad E(Y) = E(a+bX) = a + b \cdot E(X)$$

$$V(Y) = V(a+bX) = b^2 \cdot V(X) = b^2 \cdot \sigma_x^2$$

(2) Betrachten wir eine spezielle lineare Transformation, bei der $a = -\mu/\sigma_x$ und $b = 1/\sigma_x$ ist, so gilt demgemäß:

$$(7\text{-}8) \qquad z_i = \frac{x_i - \mu}{\sigma_x}$$

Aus dieser Transformation folgt:

$$(7\text{-}9) \qquad E(Z) = \frac{1}{\sigma_x} E(X) - \frac{\mu}{\sigma_x} \quad , \quad E(X) = \mu$$

$$= 0$$

$$V(Z) = \frac{1}{\sigma_x^2} \cdot V(X) \qquad \text{unter Anwendung von (7-7)}$$

$$\sigma_z^2 = 1 \quad , \quad V(Z) = \sigma_z^2$$

Das bedeutet, daß man jede normalverteilte Zufallsvariable mit den Parametern μ und σ in eine normalverteilte Zufallsvariable mit dem Erwartungswert 0 und der Standardabweichung 1 transformieren kann (vergl. die Ausführungen zur Standardisierung von Variablen in Teil I., Kap. 4.2.4). Dieses Ergebnis können wir mit der oben erwähnten Eigenschaft verbinden, daß jede Dichtefunktion f(x) einer Normalverteilung (unabhängig von der spezifischen Parameterkonstellation) zwischen den Ordinaten konstant gehaltener Intervalle $[c_1\sigma < x \leq c_2\sigma]$ konstante Flächenanteile umfaßt. Daraus folgt, daß es nicht nötig ist, für jede einzelne Parameterkonstellation (μ; σ) die Verteilungsfunktion F(x) normalverteil-

ter Zufallsvariablen zu ermitteln. Es genügt, wenn man sie einmal für die sog. **Standardnormalverteilung** $Z \sim N(0,1)$ berechnet hat: Die Verteilungsfunktion der Standardnormalverteilung wird durch das Symbol $\Phi(z)$ gekennzeichnet:

$$P(X \leq z) = \int_{-\infty}^{z} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \right) dx = \Phi(z)$$

Die Anwendung der Standardnormalverteilung wird erleichtert, indem man die Symmetrieeigenschaft ausnutzt. Wir wollen dies an einem Beispiel erläutern:

Nehmen wir an, wir hätten die Grundgesamtheit der Berliner Arbeiterhaushalte von 1900 in Haushalte »gelernter« und »ungelernter« Arbeiter aufgeteilt und aus jeder der beiden Teilpopulationen eine Stichprobe von $n = 500$ Fällen nach dem Zufallsprinzip ausgewählt. In der Gruppe der gelernten Arbeiter betrage das durchschnittliche Haushaltseinkommen im Monat $\mu_g = 1100$ RM bei einer Standardabweichung von $\sigma_g = 100$ RM. Bei den ungelernten Arbeitern seien die entsprechenden Parameter mit $\mu_u = 900$ RM und $\sigma_u = 150$ RM gegeben. Wir wollen weiterhin annehmen, die Haushaltseinkommen seien in den beiden Teilpopulationen normalverteilt (eine sicherlich unrealistische, für didaktische Zwecke aber erlaubte Annahme). Die entsprechenden Häufigkeitsdichten für die beiden Teilpopulationen sind in Abb. 7.6 dargestellt.

Wir fragen zunächst: Wie groß ist die Wahrscheinlichkeit, bei Ziehung des ersten Falles aus der Grundgesamtheit der gelernten Arbeiter einen Haushalt mit einem Monatseinkommen zwischen 1000 und 1150 RM zu erhalten? Um diese Frage zu beantworten, müßten wir »normalerweise« die Dichtefunktion (7-6) zweimal integrieren: einmal von $x = -\infty$ bis $x = 1150$, ein zweites Mal von $x = -\infty$ bis $x = 1000$. Sodann wäre die Differenz der beiden Integrale zu bilden (siehe die doppelt schraffierte Fläche in Abb. 7.7). Wir können uns die Arbeit erheblich vereinfachen, wenn wir die Einkommensvariable X standardisieren, also auch die Intervallgrenzen [$a = 1000$; $b = 1150$] in Einheiten der Standardabweichung ausdrücken:

$$(7\text{-}10) \qquad z(a) = \frac{1000-1100}{100} = -1$$

$$z(b) = \frac{1150-1100}{100} = 0{,}5$$

Die untere Intervallgrenze liegt eine Standardabweichung unterhalb, die obere Intervallgrenze eine halbe Standardabweichung oberhalb des Durchschnittseinkommens von 1100 RM. Nach der Skalentransformation (Standardisierung) wird die Reichsmark also nicht mehr in Einheiten zu 100 Pfennigen gerechnet, sondern in Einheiten der Standardabweichung, wobei eine Standardabweichung in diesem Falle den Betrag von $100 \cdot 100$ Pfennigen gleich 100 RM ausmacht. Wir können natürlich jederzeit die Beträge der neuen Skala in die ursprüngliche RM-Skala zurückrechnen. Die neue Skala aber hat den Vorteil, daß nun die Standardabweichung der Verteilung gleich 1 und das arithmetische Mittel gleich 0 ist. Für diese »Standardnormalverteilung« ist, wie bereits erwähnt, die (kumulierte) Verteilungsfunktion $\Phi(z)$ in Tabellenform in fast allen Statistik-Lehrbüchern vorhanden (siehe Anhang, Tabellen 1a und 1b). Wegen der Symmetrieeigenschaften tabelliert man die Dichtefunktion $f(z)$ nur für $z > 0$. Die Verteilungsfunktion ist, wie wir in Abschnitt 6.5 erläutert haben, das Integral der Dichtefunktion, gibt also die Fläche bzw. Flächensegmente unterhalb der Kurve der Dichtefunktion an. Wie Abbildung (7.7) verdeutlicht, ist

$$(7\text{-}11) \qquad \Phi(z_{-a}) = \int_{-\infty}^{-a} f(z)dz = 1 - \int_{-\infty}^{+a} f(z)dz = 1 - \Phi(z_{a})$$

Um unsere Frage zu beantworten, müssen wir nur den Flächenanteil bestimmen, der von den Ordinaten bei $z_a = -1$ und $z_b = 0{,}5$ eingeschlossen wird. Wir benötigen also die Differenz $\Phi(z=0{,}5) - \Phi(z=-1)$. Um die Rechnung durchführen zu können, konsultieren wir im Anhang (Tab. 1a) die Tafel der Verteilungsfunktion $\Phi(z)$[4]. Innerhalb der Tabelle findet man

[4] Aus praktischen Gründen, die erst in Kap. 8 deutlich werden, stellen manche Autoren die tabellierte Verteilungsfunktionen $\Phi(z)$ so dar, als habe man die Dichtefunktion, bildlich gesprochen, nicht von links nach

die Werte der Verteilungsfunktion, die verschiedenen positiven z-Werten zugeordnet sind. Es sind, wie oben dargelegt, die Integrale von $-\infty$ bis z der Dichtefunktion der Standardnormalverteilung. Die Werte $\Phi(z)$ für $z < 0$ ergeben sich nach Gleichung (7-11) aus der Symmetrieeigenschaft der Normalverteilung.

$$(7\text{-}12)\quad \Phi(z=0,5) = 0,6915 = 1 - 0,3085$$

$$\Phi(z=-1) = 1 - \Phi(z=1) = 1 - 0,8413 = 0,1587$$

Somit ist der gesuchte Flächenanteil

$$(7\text{-}13)\quad \Phi(z=0,5) - \Phi(z=-1) = 0,6915 - 0,1587 = 0,5328$$

oder 53,28 %. Daraus folgt, daß wir bei jeder[5] Ziehung aus der Teilpopulation der gelernten Arbeiter mit einer Wahrscheinlichkeit von $p = 0,5328$ damit rechnen können, einen Haushalt mit einem Monatseinkommen zwischen 1000 und 1150 RM zu erhalten. Für die Summe (S) der Haushalte in dieser Einkommenskategorie ergibt sich bei 500 Ziehungen der Erwartungswert $E(S) = n \cdot p = 500 \cdot 0,5328 = 266,4$ (siehe (7-4)). Wir erwarten also, daß ungefähr 266 Haushalte unserer Stichprobe in diese Einkommensklasse fallen.

Wir können nun auch die Frage beantworten, wie groß die Wahrscheinlichkeit p(A) ist, bei einer ersten Ziehung aus der Teilpopulation der ungelernten Arbeiter ebenfalls einen Haushalt mit einem Monatseinkommen zwischen 1000 und 1150 RM zu erhalten (bei $\mu_u = 900$ und $\sigma_u = 150$). Die entsprechende Ausarbeitung ist dem Leser überlassen. Sie sollte zu dem Ergebnis $p(A) = 0,205$ führen.

---

rechts, sondern von rechts nach links integriert. Dieser Art der Darstellung folgt Tab. 1b.

[5] Da wir nach jeder Ziehung den »Fall« nicht wieder in die Population »zurücklegen«, verändern sich nach jeder Ziehung die Wahrscheinlichkeiten geringfügig, weil sich die Zusammensetzung der Population verändert. Folglich sind die Stichprobenvariablen nicht völlig unabhängig voneinander. Wenn der Umfang N der Population aber sehr groß ist im Vergleich zum Stichprobenumfang n, kann man diese Ungenauigkeit vernachlässigen. Eine Daumenregel besagt, daß der Auswahlsatz $n/N < 0,05$ sein sollte. Andernfalls werden die Standardabweichungen der Zufallsvariablen mit einem Korrekturfaktor versehen; in der Regel ist das die Größe $(N-n)/(N-1)$. Zur Problematik des »Small-Population Sampling« siehe Wonnacott/Wonnacott (1972, S. 135 ff.).

Wie reagiert ein Forscher, wenn in seiner Stichprobe für beliebige Werteintervalle [$x_1 = a$, $x_2 = b$] Häufigkeiten auftreten, die »stark« von den Erwartungswerten abweichen, die er aus der Hypothese der Normalverteilung in der Population in Verbindung mit dem Modell der Zufallsauswahl abgeleitet hat? Grundsätzlich hat er zwei Möglichkeiten: Er kann die Hypothese der Normalverteilung beibehalten und annehmen, daß ihm keine »saubere«, zufallsbestimmte Stichprobenziehung gelungen ist. Er kann aber auch annehmen, daß seine Stichprobenziehung in Ordnung ist, daß dagegen die Hypothese der Normalverteilung falsch ist. Freilich werden die Stichprobenergebnisse stets »ein wenig« von den Erwartungswerten abweichen. Man braucht also Kriterien für »tolerierbare« Abweichungen, die nicht zu einer Revision der theoretischen Modelle (hier: Normalverteilung und Zufallsauswahl) zwingen. Solche Kriterien werden wir im nächsten Kapitel kennenlernen.

Wie wir dort noch sehen werden, fragt man häufig nach der Wahrscheinlichkeit, daß eine Zufallsvariable X Realisierungen im Bereich $[(\mu - c) \leqq X \leqq (\mu + c)]$ erhält (wobei c eine beliebige positive Zahl ist). Diese Wertebereiche werden als **zentrale Schwankungsintervalle** bezeichnet.

(3) Eine weitere wichtige formale Eigenschaft der Normalverteilung ist darin zu sehen, daß sie nicht nur bei einfachen Lineartransformationen erhalten bleibt, sondern auch bei der (evtl. gewichteten) Summenbildung **unabhängiger**, normalverteilter Zufallsvariablen $X_1, \ldots X_n$, wobei $X_i \sim N(\mu_i, \sigma_i)$, $i = 1,2, \ldots, n$:

$$(7\text{-}14) \qquad \sum_{i=1}^{n} b_i X_i \sim N\left(\sum b_i \mu_i, \left|\sqrt{\Sigma b_i^2 \sigma^2}\right|\right) \quad ,$$

$$\sum b_i > 0 \quad \text{und} \quad X_i \sim N(\mu_i, \sigma_i)$$

So ist das arithmetische Mittel als Stichprobenfunktion die Summe der gewichteten, identisch verteilten Stichprobenvariablen, wobei das Gewicht b für alle i ($i = 1,2, \ldots, n$) den Wert $1/n$ annimmt. Diese Stichprobenvariablen X_i sind, wie bereits erläutert, normalverteilt, wenn X in der Population normalverteilt ist. In diesem Falle ist also auch das arithmetische Mittel $\overline{X}$ normalverteilt mit dem Erwartungswert

$$(7\text{-}15) \qquad E(\bar{X}) = E(\tfrac{1}{n} X_1 + \tfrac{1}{n} X_2 + \ldots + - \tfrac{1}{n} X_n)$$

$$= \tfrac{1}{n} [E(X_1+\ldots+X_n)]$$

$$= \tfrac{1}{n} [\mu_1+\ldots+\mu_n] \quad , \quad \mu_1=\mu_2=\ldots=\mu_n=\mu$$

$$= \tfrac{1}{n} \, n \cdot \mu$$

$$= \mu$$

und der Varianz

$$(7\text{-}16) \qquad V(\bar{X}) = V(\tfrac{1}{n} X_1 + \tfrac{1}{n} X_2 + \ldots + \tfrac{1}{n} X_n)$$

$$= \tfrac{1}{n^2} V(X_1) + \ldots + \tfrac{1}{n^2} V(X_n) \quad , \text{ siehe } (7\text{-}7)$$

$$= \tfrac{1}{n^2} n \, \sigma_x^2 \quad , \quad V(X) = \sigma_x^2$$

$$= \tfrac{1}{n} \sigma_x^2 = \sigma_{\bar{x}}^2$$

Falls die Grundgesamtheit, aus der die Stichprobe gezogen wurde, einen so geringen Umfang hat, daß der Quotient n/N > 0,05 ist, wird die Standardabweichung mit einer Korrekturformel berechnet (siehe oben Fn 5):

$$(7\text{-}16') \qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Im Folgenden verzichten wir auf die Angabe dieser »Endlichkeitskorrektur«, d.h. wir gehen davon aus, daß die Grundgesamtheit im Verhältnis zur Stichprobe sehr groß ist.

Die Stichprobenfunktion X hat also eine deutlich geringere Streuung als die (einzelnen) Stichprobenvariablen, was dem intuitiven Verständnis entspricht. Die Streuung wird um so geringer, je größer der Stichprobenumfang n.

Es läßt sich zeigen, daß das arithmetische Mittel $\overline{X}$ als Stichprobenfunktion nicht nur dann normalverteilt ist, wenn die Variable X in der Population normalverteilt ist, sondern näherungsweise immer dann, wenn der Stichprobenumfang »genügend groß« ($30 \leqq n$) ist und bestimmte andere Voraussetzungen erfüllt sind, die der **Zentrale Grenzwertsatz** angibt:

Es seien $X_1, X_2, \ldots, X_n$ stochastisch unabhängige Zufallsvariablen, die sämtlich dieselbe Verteilung (nicht notwendigerweise die Normalverteilung) mit $E(X_i) = \mu$ und $V(X_i) = \sigma^2$ besitzen. Dann ist für hinreichend großes n das arithmetische Mittel $\overline{X} = (X_1 + X_2 + \ldots X_n)/n$ annähernd normalverteilt mit dem Mittelwert μ und der Standardabweichung $\sigma_x / \sqrt{n}$.
Der zentrale Grenzwertsatz wird meistens im Hinblick auf das arithmetische Mittel formuliert. Er gilt darüber hinaus auch für andere Zufallsvariablen X, die man sich additiv aus vielen unabhängigen Zufallsvariablen zusammengesetzt denken kann. Dabei muß aber sichergestellt sein, daß jede Zufallsvariable X_i nur einen kleinen Beitrag zur Summe X liefert; es darf also nicht eine einzelne Zufallsvariable X_i die Summe X dominieren.

Daß das arithmetische Mittel unabhängig von der Verteilungsform der Stichprobenvariablen gegen die Normalverteilung konvergiert, wird anhand des Würfelbeispiels anschaulich. Bei $n = 1$ (einem einzigen Wurf) ergibt sich für die einzelnen Ereignismöglichkeiten eine Gleichverteilung (siehe die folgende Abbildung). Schon bei $n = 3$ nimmt die Wahrscheinlichkeitsfunktion für den Mittelwert eine glockenförmige Gestalt an (siehe Abb. 7.8).

Ein ähnliches Entwicklungsmuster stellt sich auch bei stetigen Variablen ein.

7.4 Mit der Normalverteilung verbundene Verteilungsmodelle

7.4.1 Die Chi-Quadrat-Verteilung

Eine zweite Stichprobenfunktion, mit der wir es neben dem arithmetischen Mittel häufig zu tun haben, ist die Varianz

$$(7\text{-}17) \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Statistiker haben herausgefunden, daß es günstiger ist, die in der Regel

unbekannte Varianz σ^2 der Population mit einer Stichprobenstreuung zu schätzen, in deren Nenner nicht der Stichprobenumfang n, sondern die Zahl der sog. Freiheitsgrade, n-1, steht. Den Begriff der **Freiheitsgrade** (df) kann man sich wie folgt veranschaulichen: Um s^2 zu bestimmen, muß man zuvor das arithmetische Mittel $\bar{x}$ berechnen. Das bedeutet aber, daß von den n Differenzen $(x_i - \bar{x})$ nur (n-1) unabhängig sind, also beliebige Werte aus dem Stichprobenraum realisieren können. Wenn erst einmal (n-1) Differenzen vorliegen, ist die n-te Differenz determiniert, da ja die Summe aller Differenzen Null ergeben muß.

Die Varianz s^2 streut von Stichprobe zu Stichprobe[6] in rechtsschiefer Verteilung um die Populationsvarianz σ^2. Wenn X normalverteilt ist, ist die Verteilung von s^2 proportional zu der sog. χ^2-(Chi-Quadrat-)Verteilung. Ihre Kurve weist für jede Kombination von σ^2 und n eine jeweils andere Verlaufsform auf. Wir erinnern uns: auch die Normalverteilungskurve variierte mit den Parametern μ und σ. Die Vielfalt der Normalverteilungskurven konnte durch die Transformation der Variablen in z-Werte (siehe (7-8) zu einer einzigen reduziert werden. Eine ähnliche (allerdings weniger weitgehende) Lösung gelingt hinsichtlich der Verteilung von s^2, wenn man die folgende Transformation durchführt:

$$(7\text{-}18) \qquad \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Die Verteilungsfunktionen für χ^2 liegen ebenfalls in tabellierter Form vor (siehe Anhang A, Tab.3). Anders als bei Z bleibt die Verteilung von χ^2 von dem jeweiligen Stichprobenumfang (bzw. den Freiheitsgraden) abhängig, wie die Abbildung 7.9 verschiedener Dichtefunktionen zeigt. (Um die Zahl der Freiheitsgrade anzuzeigen, fügt man statt »df« auch häufig das Subskript $\nu = n-1$ dem Symbol χ^2 hinzu).

Die Chi-Quadrat-Verteilungen haben folgende Mittelwerte und Varianzen:

$$(7\text{-}19) \qquad E(\chi_{n-1}^2) = n-1$$

$$V(\chi_{n-1}^2) = 2(n-1)$$

Mit größer werdendem n nähert sich die Chi-Quadrat-Verteilung einer

[6] Wir führen einstweilen das Kürzel »s^2« weiter, differenzieren aber später die Symbolik, um eindeutiger zwischen der (deskriptiven) Stichprobenkennzahl und der Schätzgröße für die entsprechende Kennzahl der Populationsverteilung zu unterscheiden (siehe auch Abschn. 7.4.2).

Normalverteilung mit dem Mittelwert $\mu = n-1$ und der Standardabweichung $\sigma = \sqrt{2(n-1)}$.

Im Rahmen dieses Kurses werden wir die Chi-Quadrat-Verteilung vor allem anwenden, um die Hypothese eines fehlenden Zusammenhangs zwischen zwei nominal skalierten Variablen anhand von Stichprobendaten zu testen (siehe Kap. 8.7.2). Wir hatten ja als Maß für einen solchen Zusammenhang eine Kennzahl

$$\chi^2 = \sum_{k=1}^{K} \frac{(f_{b_k} - f_{e_k})^2}{f_{e_k}}$$

definiert (siehe Gleichung (4-7) in Teil I), die leider mit dem gleichen Symbol abgekürzt wird, wie die transformierte Stichprobenfunktion (7-18). Die Maßzahl χ^2 ähnelt formal sehr stark dem Varianzausdruck. Es muß deshalb nicht überraschen (auch wenn die identischen Symbole für etwas Verwirrung sorgen können), daß die Maßzahl χ^2 in guter Annäherung χ^2 verteilt ist.

Die (allgemeine) Chi-Quadrat-Verteilung läßt sich aber noch für eine Reihe weiterer Zwecke anwenden, z. B. dann, wenn sog. Vertrauensintervalle (siehe Kap. 8) für die mit s^2 geschätzte Populationsvarianz σ^2 zu ermitteln sind.

Betrachten wir noch einmal den Ausdruck (7-18). Offensichtlich läßt er sich wegen (7-17) umformen zu

$$\chi_{n-1}^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\sigma^2} = \frac{1}{\sigma^2} \Sigma (x_i-\bar{x})^2 \qquad (7\text{-}20)$$

χ^2 ist also eine gewichtete Summenvariable. Das bringt uns zu der allgemeineren Bestimmung der Chi-Quadrat-Verteilung: Sind $X_1, \ldots, X_n$ unabhängige, standardnormalverteilte (oder entsprechend transformierte) Zufallsvariablen (wir wissen bereits, daß dann auch ihre Lineartransformationen normalverteilt sind), so hat die Summe S_n der quadrierten X_i, $(i = 1,2, \ldots, n)$

$$S_n = X_1^2 + \ldots + X_n^2$$

eine Dichtefunktion, die als Chi-Quadrat-Verteilung mit n Freiheitsgraden bezeichnet wird[7]. Bei nur einem Freiheitsgrad (wenn also die S_n nur aus einem Glied besteht), ist die Chi-Quadrat-Verteilung gleich der quadrierten Standardnormalverteilung.

Wir hatten gesehen, daß auf Grund des Zentralen Grenzwertsatzes das arithmetische Mittel immer ab einer Stichprobengröße $n > 30$ annähernd normalverteilt ist. Die Verteilungsform von X in der Population kann also ignoriert werden. Im Hinblick auf die Stichprobenfunktion s^2 und deren Chi-Quadrat-Verteilung muß aber an der Normalverteilungsvoraussetzung der Ursprungsvariable X in der Population festgehalten werden (Kendall/Stuart (1986), Band I, S.537). Eine Verletzung dieser Annahme ist aber um so weniger gravierend, je größer n.

7.4.2 Die t-Verteilung

Wir hatten in Abschnitt 7.3 festgestellt, daß das arithmetische Mittel $\overline{X}$ als Stichprobenfunktion normalverteilt ist mit der Standardabweichung $\sigma_x / \sqrt{n}$. Die Standardabweichung σ_x der Variable X in der Population ist aber in der Regel gar nicht bekannt. In diesem Falle müssen wir die in der Population gegebene und mit »σ« bezeichnete Standardabweichung durch die in der Stichprobe ermittelte und mit »s« abgekürzte Standardabweichung schätzen. Um als Schätzer für die Standardabweichung σ in der Population dienen zu können, muß, wie bereits erwähnt, die Formel für die Stichprobenvarianz bzw. für ihren Wurzelausdruck etwas modifiziert werden: aus $s^2 = 1/n \Sigma (x_i - \bar{x})^2$ wird $\hat{\sigma}^2 = 1/(n-1) \Sigma (x_i - \bar{x})^2$. Ohne diese Modifikation ist die Stichprobenvarianz kein »erwartungstreuer« Schätzer. (Was Erwartungstreue bedeutet, wird Abschn. 8.3 erläutert.) Bei großem n fällt die Korrektur des Divisors rechnerisch allerdings nicht ins Gewicht. Bisher konnten wir die Stichprobenfunktion $\overline{X}$ (ebenso wie die Stichprobenvariable X) standardisieren, indem wir ihre Differenz zum Erwartungswert $E(\overline{X}) = \mu$ durch die Standardabweichung ihrer Verteilung dividiert haben:

$$(7\text{-}21) \qquad Z = \frac{\bar{x} - \mu}{\sigma_x / \sqrt{n}}$$

Wenn wir nun σ durch die empirische Stichprobenkennzahl $\hat{\sigma}$ ersetzen, erhalten wir eine neue Größe

---

[7] Diesmal werden keine Freiheitsgrade durch Berechnung des Mittelwertes oder anderer Parameter »verbraucht«.

$$(7\text{-}22) \qquad t_{n-1} = \frac{\bar{x} - \mu}{s_x/\sqrt{n-1}} = \frac{\bar{x} - \mu}{\hat{\sigma}_x/\sqrt{n}}$$

Während σ eine Konstante bezeichnete, nämlich die in der Population gegebene (gleichwohl unbekannte) Standardabweichung von X, so bezeichnet $\hat{\sigma}$ nun eine Zufallsgröße: als Schätzer von σ kann sie, worüber wir schon im vorigen Abschnitt gesprochen haben, mehr oder weniger stark von der Zielgröße σ abweichen. (Man setzt das Symbol ^ über den Populationsparameter, um einen Schätzer als solchen zu kennzeichnen.) Das bedeutet, daß im Vergleich zu den standardnormalverteilten Z-Werten des arithmetischen Mittels nun ein weiterer Unsicherheitsfaktor eingeführt wird. Diese Unsicherheit ist um so größer, je kleiner die Stichprobe ist. Bei größeren Stichproben kann man davon ausgehen, daß $\hat{\sigma}$ weniger stark um σ schwankt als bei kleineren Stichproben. Gegenüber der Normalverteilung von Z hat die Verteilung von t (man spricht auch nach dem Pseudonym ihres Entdeckers, W. Grosset, von »Student's t«) folglich eine größere Spannbreite.

Um keine Mißverständnisse aufkommen zu lassen: Nicht die Stichprobenfunktion $\bar{X}$ streut stärker, sondern unsere **Einschätzung** dieser Streuung wird unsicherer, wenn σ nicht bekannt ist, sondern durch s bzw. $\hat{\sigma}$ geschätzt werden muß. Diese Unsicherheit müssen wir, insbesondere beim Testen von Hypothesen (siehe Kap. 8) durch breitere Schwankungsintervalle (sog. Konfidenzintervalle), d. h. durch ein etwas anderes Verteilungs**modell** berücksichtigen.

Mit wachsendem n nähert sich die t-Verteilung immer stärker der Normalverteilung an. Daraus folgt, daß es nicht eine »Standard«-t-Verteilung geben kann, sondern für jeden (kleinen) Stichprobenumfang eine besondere t-Verteilung. In der Literatur werden als Faustregel für hinlänglich gute Annäherung Mindestgrößen von $n = 25$ bis $n = 100$ genannt. Die t-Verteilung ist aber nicht nach dem Stichprobenumfang n, sondern nach den Freiheitsgraden n-1 tabelliert (siehe hierzu auch Abschn. 7.4.1). Die Abbildung 7.10 zeigt, wie sich die t-Verteilung relativ rasch an die Normalverteilung anpaßt.

Bei der Ableitung der t-Verteilung wird vorausgesetzt, daß die entsprechende Variable X in der Population normalverteilt ist. Wenn wir über eine genügend große Stichprobe verfügen, können wir aber auch bei unbekanntem, durch $\hat{\sigma}$ geschätztem σ die Normalverteilung statt der t-Verteilung benutzen. Selbst bei relativ kleinen Stichproben gilt das t-Verteilungsmodell als ziemlich »robust«, d. h., es bleibt auch dann ein adäquates Verteilungsmodell für bestimmte statistische Maßzahlen, wenn die Voraussetzung der Normalverteilung von X nicht erfüllt ist (Bortz 1979, S.166).

Nicht nur das standardisierte arithmetische Mittel ist t-verteilt. Allgemein gilt: Sind X und Y zwei unabhängig voneinander verteilte Zufallsvariablen und ist X standardnormalverteilt, Y hingegen χ^2-verteilt mit k Freiheitsgraden, so hat der Quotient

$$(7\text{-}23) \qquad t_k = \frac{x}{\sqrt{y/k}},$$

eine Dichtefunktion, die der t-Verteilung mit k Freiheitsgraden entspricht.

7.4.3 Die F-Verteilung

Wenn X_m und X_n zwei voneinander unabhängige χ^2-verteilte Variablen mit m bzw. n Freiheitsgraden sind, hat der Quotient

$$(7\text{-}24) \qquad F_{m,n} = \frac{\frac{x}{m}}{\frac{y}{n}} = \frac{x \cdot n}{y \cdot m}$$

eine Dichtefunktion, die als F-Verteilung bezeichnet wird. Sie ist sowohl von den Freiheitsgraden der ersten als auch von den Freiheitsgraden der zweiten Variablen abhängig (siehe Abb. 7.11). Die Freiheitsgrade werden als Index von F notiert, wobei man üblicherweise die Zählerfreiheitsgrade zuerst nennt. Abb. 7.11 zeigt zwei F-Verteilungen für unterschiedliche Freiheitsgrade. Man beachte, daß bei der F-Verteilung im Unterschied zur Z- oder t-Verteilung nur positive Werte vorkommen. Da, wie wir eben sahen, Varianzen χ^2-verteilt sind, ist der Quotient zweier Varianzen bzw. Varianzschätzer F-verteilt. So läßt sich dieses Verteilungsmodell z. B. heranziehen, wenn man testen will, ob in der Population ein Zusammenhang zwischen einer nominal-skalierten Variablen X und einer metrisch-skalierten Variablen Y besteht (siehe Teil I, Kap. 4.2.5). Auch im Rahmen der Regressionsanalyse kann der F-Test angewandt werden. Seine Anwendung ist aber auch bei größeren Stichproben an die Voraussetzung gebunden, daß die Stichprobenvariablen normalverteilt sind.

Zwischen der t-Verteilung und der F-Verteilung besteht folgende Beziehung:

(7-25) $t_n^2 = F_{1,n}$

Die quadrierte t-Verteilung mit n Freiheitsgraden ist identisch mit der F-Verteilung für 1 Zählerfreiheitsgrad und n Nennerfreiheitsgrade. Näheres über die Beziehungen der verschiedenen Verteilungsmodelle zueinander findet man in Hays (1973, S.451ff.).

Abb. 7.1: Fiktive Einkommensverteilung

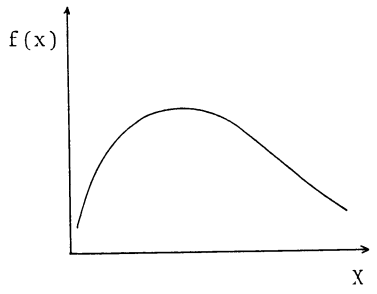
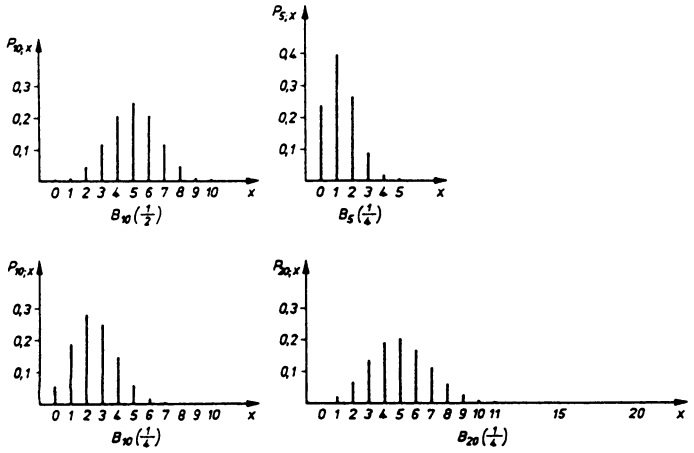


Abb. 7.2: Entwicklung der Binomialverteilung für die Anzahl S der Zentrumsabgeordneten bei 4 Ziehungen mit P(Z)=0,2; Z ≙ X=1; Z ≙ X=0; $S = \Sigma X_i$

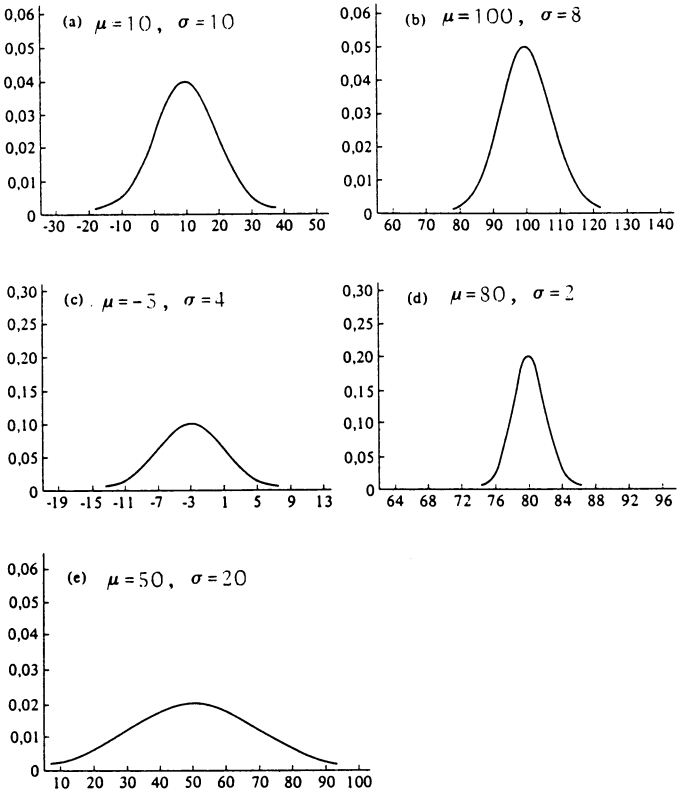
Folge von Einzelereign.	Wahrscheinlichk. f. Ereignisfolge	Wert der Stichprobenfkt. S	Wahrscheinlichk. f. d. Wert S=s
$\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3 \cap \bar{Z}_4$	$(1-0,2)^4$	S = 0	$(1-0,2)^4$
$Z_1 \cap \bar{Z}_2 \cap \bar{Z}_3 \cap \bar{Z}_4$	$0,2^1 \cdot (1-0,2)^3$	S = 1	$4 \cdot 0,2 \cdot (1-0,2)^3$
$\bar{Z}_1 \cap Z_2 \cap \bar{Z}_3 \cap \bar{Z}_4$	$0,2^1 \cdot (1-0,2)^3$		
$\bar{Z}_1 \cap \bar{Z}_2 \cap Z_3 \cap \bar{Z}_4$	$0,2^1 \cdot (1-0,2)^3$		
$\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3 \cap Z_4$	$0,2^1 \cdot (1-0,2)^3$		
$Z_1 \cap Z_2 \cap \bar{Z}_3 \cap \bar{Z}_4$	$0,2^2 \cdot (1-0,2)^2$	S = 2	$6 \cdot 0,2^2 \cdot (1-0,2)^2$
⋮	⋮		
$\bar{Z}_1 \cap \bar{Z}_2 \cap Z_3 \cap Z_4$	$0,2^2 \cdot (1-0,2)^2$		
$Z_1 \cap Z_2 \cap Z_3 \cap \bar{Z}_4$	$0,2^3 \cdot (1-0,2)^1$	S = 3	$4 \cdot 0,2^3 \cdot (1-0,2)^1$
$Z_1 \cap Z_2 \cap \bar{Z}_3 \cap Z_4$	$0,2^3 \cdot (1-0,2)^1$		
$Z_1 \cap \bar{Z}_2 \cap Z_3 \cap Z_4$	$0,2^3 \cdot (1-0,2)^1$		
$\bar{Z}_1 \cap Z_2 \cap Z_3 \cap Z_4$	$0,2^3 \cdot (1-0,2)^1$		
$Z_1 \cap Z_2 \cap Z_3 \cap Z_4$	$0,2^4$	S = 4	$0,2^4$

Abb. 7.3: Binomialverteilungen mit unterschiedlichen n (erster Index von B) und p (zweiter Index)



Quelle: Claus/Ebner 1977, S. 145

Abb. 7.4: Verschiedene Normalverteilungen



Quelle: Bortz 1979, S. 93 (Notation geändert)

Abb. 7.5: Dichtefunktion der Normalverteilung

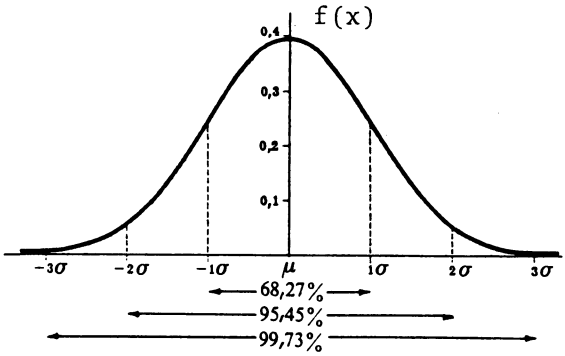
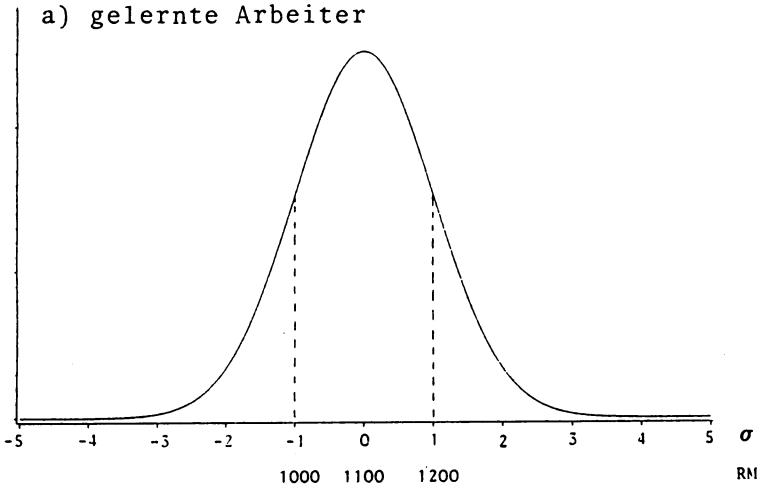


Abb. 7.6: Fiktive Einkommensverteilungen



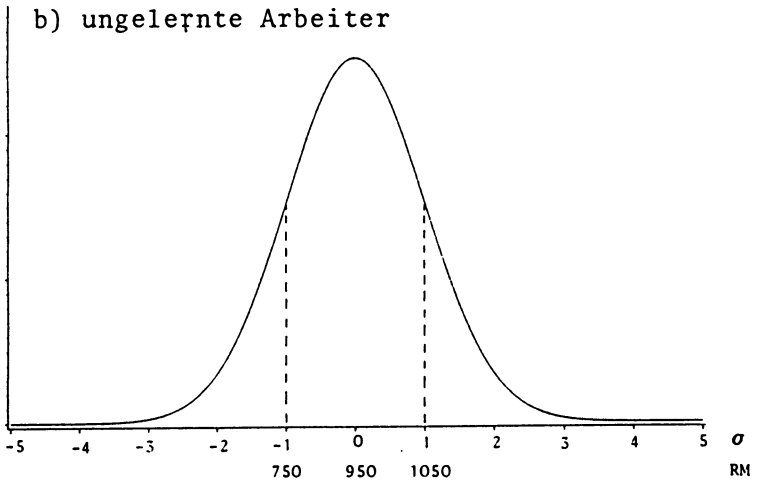


Abb. 7.7: Veranschaulichung zu Gleichung (7-11)

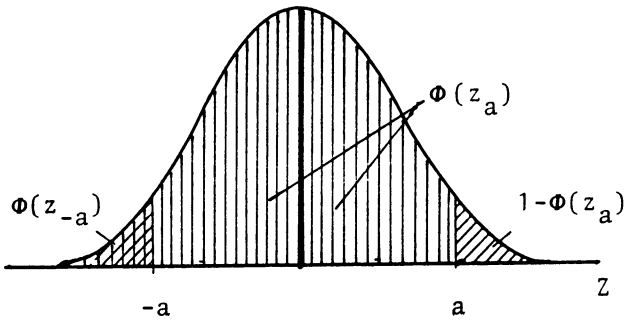
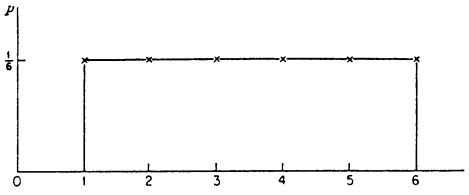
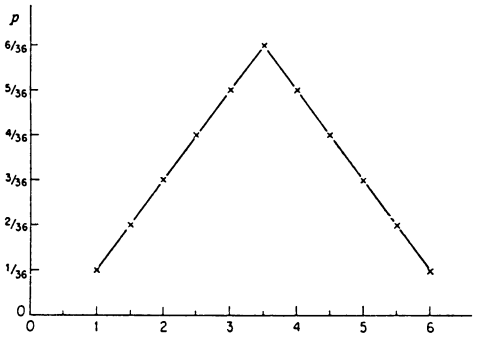


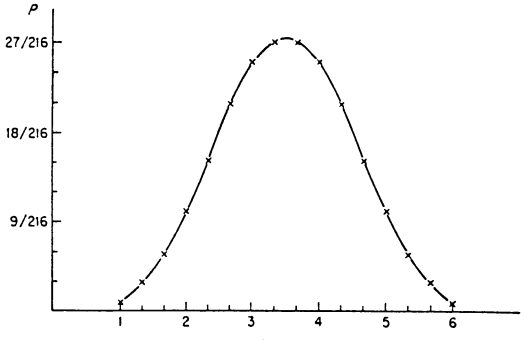
Abb. 7.8: Annäherung der Verteilung der Stichprobenmittel an das Modell der Normalverteilung



Theoretische Verteilung der Wahrscheinlichkeiten für unterschiedliche Würfelergebnisse (1 Wurf mit perfektem Würfel)



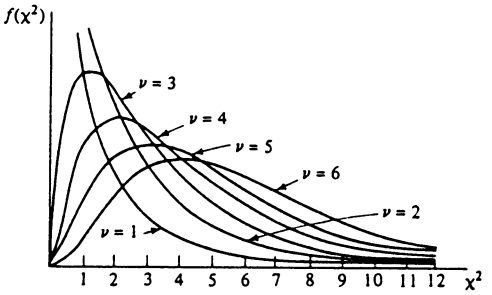
Theoretische Stichprobenverteilung des arithmetischen Mittels der Augenzahlen bei 2 Würfen mit perfektem Würfel



Theoretische Stichprobenverteilung des arithmetischen Mittels der Augenzahlen bei 3 Würfen mit perfektem Würfel

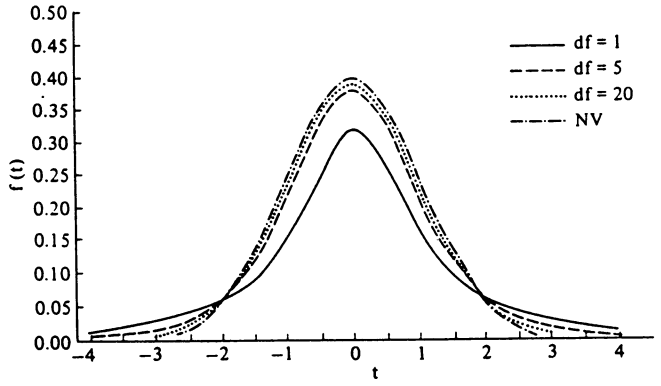
Quelle: Blalock 1960, S. 139 f.

Abb. 7.9: Verschiedene Dichtefunktionen für Chi-Quadrat mit unterschiedlichen Freiheitsgraden



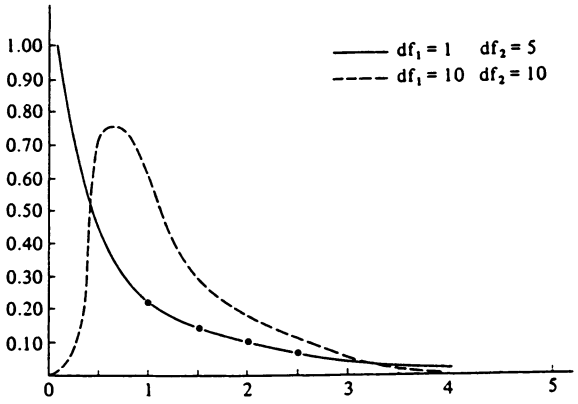
Quelle: Kmenta 1971, S. 141

Abb. 7.10: Annäherung der t-Verteilung an die Normalverteilung



Quelle: Bortz 1979, S. 104

Abb. 7.11: F-Verteilungen (Dichtefunktionen) mit unterschiedlichen Zählerfreiheitsgraden (df_1) und Nennerfreiheitsgraden (df_2)



Quelle: Bortz 1979, S. 105

KAPITEL 8

Schätzen und Testen

Wir haben jetzt schon mehrmals folgende Problemstellung berührt: Wie können wir aus Merkmalen (statistischen Kennwerten) einer Stichprobenverteilung auf Merkmale der Populationsverteilung schließen (»Repräsentationsschluß«), und wie können wir Hypothesen über Merkmale der Populationsverteilung anhand von Stichprobendaten testen (»Inklusionsschluß«)? In den beiden vorangegangenen Kapiteln haben wir die Prinzipien erörtert, die es uns erlauben, diese Fragen zu beantworten. Jetzt wollen wir sie anwenden:

8.1 Erstes Beispiel: Intervallschätzung des arithmetischen Mittels

Aus der Grundgesamtheit der Reichstagsabgeordneten kennen wir u. a. die Verteilung von X = Dauer der Mitgliedschaft im Parlament. Sie hat das arithmetische Mittel $\mu = 7{,}16$ und die Standardabweichung $\sigma = 6{,}332$. Aus dieser Grundgesamtheit ziehen wir nun eine Zufalls-Stichprobe von $n = 100$[1] Fällen mit dem arithmetischen Mittel $\bar{x} = 6{,}30$ und der Standardabweichung $s = 5{,}59$. Wie würden wir vorgehen, wenn uns das μ der Grundgesamtheit nicht bekannt wäre und wir es anhand dieses Stichprobenergebnisses schätzen wollten? Es scheint naheliegend, als »Schätzer« von μ das Stichprobenmittel $\bar{x}$ zu verwenden. (In Abschnitt 8.3 werden wir sehen, daß das nicht so selbstverständlich ist.) Unser Stichprobenmittel weicht vom Populationsmittel ab; damit ist immer zu rechnen. Deshalb erscheint es sinnvoll, nicht nur den sog. Punktwert für μ zu schätzen (»Punktschätzer«), sondern ein Intervall, das mit einer angebbaren Wahrscheinlichkeit den Parameter μ einschließt.

Wir hatten schon im vorigen Kapitel erwähnt, daß der »induktive« Schluß beim Schätzen von Parametern sich aus der Umkehrung der »deduktiven« Perspektive ergibt. Das heißt, wir bestimmen zunächst einmal die Verteilung der Stichprobenfunktion $\bar{X}$, die man erhielte, wenn »unendlich oft« Stichproben aus der gleichen Grundgesamtheit gezogen würden.

---

[1] Wir ignorieren hier die Problematik des »Small Population Sampling«, siehe Kap. 7, Fn. 5.

Bei $n = 100$ können wir uns auf den Zentralen Grenzwertsatz stützen und davon ausgehen, daß $\bar{X}$ normalverteilt ist mit dem Erwartungswert $E(\bar{X}) = \mu = 7{,}16$ und der Standardabweichung $\sigma_{\bar{x}} = \sigma_x / \sqrt{n} = 6{,}332/10 = 0{,}633$. (Man bezeichnet die Standardabweichung eines Schätzers auch als »Standardfehler«.) Unser Stichprobenergebnis liegt also um $(7{,}16-6{,}30)/0{,}633 = 1{,}35$ Standardabweichungen vom wahren Wert, μ, entfernt.

Aus der Verteilungsfunktion der Normalverteilung wissen wir (siehe Kap. 7 und Tab. 1 a,b im Anhang), daß 95 % aller Stichprobenmittel (bei unendlich oft wiederholten Stichprobenziehungen) in einem Intervall

(8-1) $$[(\mu - 1{,}96\,\sigma_{\bar{x}}) \leq \bar{X} \leq (\mu + 1{,}96\,\sigma_{\bar{x}})]$$

liegen. Oder, anders formuliert: Für ein arithmetisches Mittel $\bar{x}$ aus einer Stichprobe gilt, daß es mit einer Wahrscheinlichkeit von $p = 0{,}95$ aus dem in (8-1) genannten Intervall stammt. Durch einfache algebraische Umformung erhalten wir aus (8-1)

(8-2)

$$P\left(\mu - 1{,}96\frac{\sigma_x}{\sqrt{n}} \leq \bar{x} \leq \mu + 1{,}96\frac{\sigma_x}{\sqrt{n}}\right) = 0{,}95$$

Der Parameter μ wird subtrahiert

$$P(-1{,}96\,\sigma_{\bar{x}} \leq \bar{x} - \mu \leq 1{,}96\,\sigma_{\bar{x}}) = 0{,}95$$

Die Ungleichung wird mit (-1) multipliziert (die Ungleichheitszeichen werden folglich invertiert)

$$P(1{,}96\,\sigma_{\bar{x}} \geq \mu - \bar{x} \geq -1{,}96\,\sigma_{\bar{x}}) = 0{,}95$$

Das arithmetische Mittel $\bar{x}$ wird addiert

$$P(1{,}96\,\sigma_{\bar{x}} + \bar{x} \geq \mu \geq \bar{x} - 1{,}96\,\sigma_{\bar{x}} = 0{,}95$$

Rechte und linke Seite der Ungleichung werden vertauscht

$$P(\bar{x} - 1{,}96\sigma_{\bar{x}} \leq \mu \leq \bar{x} + 1{,}96\sigma_{\bar{x}}) = 0{,}95.$$

Die Ungleichung in der letzten Zeile von (8-2) bedeutet nicht, daß μ nun zu einer Zufallsvariable geworden wäre; μ ist nach wie vor ein festliegender Parameter der Populationsverteilung. Die Wahrscheinlichkeitsaussage bezieht sich auf $\overline{X}$, genauer auf das Intervall $\bar{x} \pm 1{,}96\,\sigma$. Wir haben es also mit einem weiteren Typ von Zufallsvariablen zu tun: Intervallen (bzw. Paaren von Intervallgrenzen), die mit $\overline{X}$ bzw. den jeweiligen Realisationen der Stichprobenvariablen $X_1, \ldots, X_n$ variieren und mit einer bestimmten Wahrscheinlichkeit den Populationsparameter einschließen. Die Abb. 8.1 veranschaulicht die Variation dieser Zufallsvariablen anhand eines anderen Beispiels (Körperlänge in Inches), das wir dem Lehrbuch von Wonnacott/Wonnacott 1972, S. 145 entnehmen.

Wenn $\bar{x}$ mit einer Wahrscheinlichkeit von 95 % in das »Schwankungsintervall« $\mu \pm 1{,}96\sigma$ fällt, dann kann man auch »umgekehrt« schließen, daß die Zufallsvariable $(\bar{x} \pm 1{,}96\sigma)$ mit 95 % Wahrscheinlichkeit den Populationsparameter μ einschließt. Man nennt dieses Intervall (95 %-)**Vertrauensintervall** oder auch »Konfidenzintervall«. In unserem Beispiel erhalten wir das 95 %-Vertrauensintervall für μ mit

$$(8\text{-}3) \qquad 6{,}30 \pm 1{,}96 \cdot 0{,}633 = 6{,}30 \pm 1{,}96 \cdot 0{,}633$$

Als untere Grenze des Vertrauensintervalls für die durchschnittliche Dauer der Mitgliedschaft im Parlament schätzen wir also 5,059 Jahre, als obere Grenze 7,541 Jahre. Da in diesem Falle $\mu = 7{,}16$ Jahre bekannt ist, wird die Schätzung als korrekt bestätigt. Ansonsten können wir nur Wahrscheinlichkeitsaussagen machen: Die Wahrscheinlichkeit, daß das 95 %-Vertrauensintervall μ **nicht** einschließt, ist $1-0{,}95 = 0{,}05$. Sie wird üblicherweise mit dem Buchstaben α bezeichnet und **Irrtumswahrscheinlichkeit** genannt. Mit Alpha sind bei der Standardnormalverteilung auch die entsprechenden z-Werte, $z_{\alpha/2}$ und $z_{1-\alpha/2}$, für die untere und die obere Intervallgrenze gesetzt. Da die Standardnormalverteilung symmetrisch ist, gilt $z_{\alpha/2} = -z_{1-\alpha/2}$. $(\alpha/2)$ und $(1-\alpha/2)$ geben Quantile an, wie sie in Teil I, Kap. 3.2 besprochen wurden, diesmal aber bezogen auf eine theoretische Verteilung in Form einer Wahrscheinlichkeitsdichtefunktion. $z_{0{,}975}$ bezeichnet also das 97,595-Quantil der Z-Verteilung. In den meisten Normalverteilungstabellen sind den halben Irrtumswahrscheinlichkeiten die $z_{1-\alpha/2}$-Werte zugeordnet. In der Tabelle 1b des Anhangs A findet man bei $p(z \geq z_p) = 0{,}025 = \alpha/2$ den Wert $z_p = z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$. Wählt man mit $\alpha = 0{,}01$ eine geringere Irrtumswahrscheinlichkeit, so findet man bei $p = \alpha/2 = 0{,}005$ den Wert $z_{1-\alpha/2} = 2{,}575$ (nach Interpolation). Folglich ist $z_{\alpha/2} = -2{,}575$ und das 99 %-Vertrauensintervall ist gegeben mit

$$(8\text{-}4) \qquad P(\bar{x} - 2{,}575\sigma \leq \mu \leq \bar{x} + 2{,}575\sigma) = 0{,}99$$

Je geringer die Irrtumswahrscheinlichkeit, die man akzeptiert, um so größer das Vertrauensintervall, um so unpräziser also die Schätzung. Je stärker man den jeweils interessierenden Parameter, hier: μ, eingrenzen will, desto höher die Irrtumswahrscheinlichkeit, die man dabei akzeptieren muß.

Spielen wir noch die Variante durch, bei der die Varianz σ_x^2 der Population nicht bekannt ist. Als Schätzer benutzen wir, wie in Abschn. 7.4.2 erläutert, die modifizierte Stichprobenvarianz $\hat{\sigma}^2 = 1/(n-1) \Sigma (x_i - \bar{x})^2 = 31{,}25$. Die geschätzte Standardabweichung des arithmetischen Mittels ist also $\hat{\sigma}_{\bar{x}} = \sqrt{31{,}25} / \sqrt{100} = 0.559$.

Obwohl wir mit $n = 100$ einen Stichprobenumfang vorliegen haben, bei dem man, gestützt auf den Zentralen Grenzwertsatz, im allgemeinen die Standardnormalverteilung heranzieht, wollen wir hier der Übung wegen die t-Verteilung (siehe Anhang A, Tab.2) zugrundelegen. Die Irrtumswahrscheinlichkeit sei erneut $\alpha = 0{,}05$. Wir bestimmen folglich das Intervall

$$(8\text{-}5) \qquad \bar{x} \pm t_{(99)0,025}\, \hat{\sigma}_{\bar{x}} = 6{,}30 \pm 1{,}98 \cdot 0{,}559$$

$$= 6{,}30 \pm 1{,}11$$

Die untere Intervallgrenze liegt somit bei 5,19, die obere bei 7,41 Jahren. Wir haben zwar mit unserer Stichprobenstreuung die Populationsvarianz unterschätzt, das Konfidenzintervall ist deshalb trotz des im Vergleich zur Z-Verteilung etwas höheren t-Wertes (1,98 statt 1,96) enger geworden, schließt aber immer noch den Parameter μ ein.

8.2 Zweites Beispiel: Test auf »Signifikanz« einer Mittelwertdifferenz

Wir wollen prüfen, ob sich in der Grundgesamtheit der Reichstagsabgeordneten adlige und nicht-adlige Abgeordnete hinsichtlich der durchschnittlichen Dauer ihrer Parlamentszugehörigkeit unterscheiden. (Mit anderen Worten, wir wollen wissen, ob ein »Zusammenhang« besteht zwischen der Variablen »Mandatsdauer der Reichstagsabgeordneten« und der Variablen »Zugehörigkeit zum Adel - ja oder nein«.) Wenn die Zugehörigkeit zum Adel keinerlei Einfluß auf die Dauer der Mitgliedschaft im Reichtstag hätte, dürften sich die beiden Mittelwerte μ_1 und μ_2 nicht unterscheiden. (Der Index »1« soll hier und im folgenden die Gruppe der Adligen, der Index »2« die Gruppe der Nichtadligen kennzeichnen.) Wir wollen dies als unsere Hypothese formulieren:

(8-6) $\quad H: \mu_1 - \mu_2 = 0$

Sie soll anhand zweier unabhängiger Zufallsstichproben mit $n_1 = n_2 = 50$ überprüft werden, die aus den (getrennten) Grundgesamtheiten der adligen und nicht-adligen Abgeordneten gezogen werden. (Auch eine einzige Stichprobe wäre möglich, solange die einzelnen Fälle unabhängig voneinander erhoben werden. Dann ließen sich die Adligen und Nicht-Adligen zu jeweils einer Teilstichprobe zusammenfassen, die von der anderen unabhängig wäre. In diesem Falle wäre mit ungleichen Stichprobenumfängen zu rechnen, wenn in der Population $N_1 \neq N_2$ ist. Um den Test durchzuführen, ist es zwar nicht erforderlich, daß die beiden Teilstichproben den gleichen Umfang aufweisen. Bei kleinen Stichproben sollten die Umfänge aber auch nicht stark differieren.)

Selbst wenn in der Grundgesamtheit tatsächlich $\mu_1 = \mu_2$ gegeben wäre, müßten wir damit rechnen, daß sich die beiden Stichprobenmittelwerte auf Grund von Zufallseinflüssen unterscheiden, die bei der Stichprobenziehung wirksam sind. Wir können also nicht unbedingt von $(\bar{x}_1 - \bar{x}_2) \neq 0$ auf einen (wenn auch vielleicht nur schwachen) Zusammenhang zwischen Zugehörigkeit/Nichtzugehörigkeit zum Adel und der Mandatsdauer schließen. Die Frage ist also, wie groß (oder klein) die Differenz $(\bar{x}_1 - \bar{x}_2)$ sein muß, damit wir uns berechtigt fühlen dürfen, die Hypothese (8-6) zu verwerfen oder beizubehalten. Die Antwort setzt voraus, daß wir die Differenz $(\bar{X}_1 - \bar{X}_2)$ als eine Zufallsvariable auffassen, die mit unterschiedlichen Wahrscheinlichkeiten unterschiedliche Werte (Wertebereiche) annehmen kann. Als erstes müssen wir überlegen, welche Verteilungsfunktion diese Zufallsvariable hat:

Wenn die Variable X in beiden Gruppen normalverteilt ist, dann sind auch die Mittelwerte $\bar{X}_1$ und $\bar{X}_2$ normalverteilt. Wenn dies nicht der Fall ist, aber $(n_1 + n_2) \geqq 50$ ist, können wir den Zentralen Grenzwertsatz in Anspruch nehmen und wiederum die (approximierte) Normalverteilung für $\bar{X}_1$ und $\bar{X}_2$ voraussetzen.

Wir hatten schon in Kap. 7 festgestellt, daß a) jede Lineartransformation einer normalverteilten Variablen und b) jede Summe unabhängiger, normalverteilter Variablen wiederum normalverteilt ist. Daraus folgt, daß auch jede Kombination $Z = aX_1 + bX_2$ normalverteilt ist, wenn X_1 und X_2 voneinander unabhängige, normalverteilte Zufallsvariablen sind. Angewandt auf die Mittelwertdifferenz in unserem Beispiel ist $a = 1$ und $b = -1$, so daß sich für Erwartungswert und Varianz folgendes ergibt:

(8-7) $$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2)$$

$$= \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{x}(1)-\bar{x}(2)} = (+1)^2 V(\bar{X}_1) + (-1)^2 V(\bar{X}_2)$$

nach (6-36) in Verbindung mit (6-39)

Die Varianz der Differenz $\bar{X}_1 - \bar{X}_2$ ist also die gleiche wie die Varianz der Summe. Daß $V(\bar{X}_1 - \bar{X}_2)$ größer ist als $V(\bar{X}_1)$ oder $V(\bar{X}_2)$ ist plausibel, wenn man bedenkt, daß in etwa der Hälfte aller Zufallsexperimente $\bar{x}_1$ und $\bar{x}_2$ von ihrem jeweiligen Erwartungswert μ_1 bzw. μ_2 in entgegengesetzter Richtung abweichen, so daß hierbei größere Absolutdifferenzen entstehen als bei $\bar{X} - \mu$. Zusammenfassend können wir also festhalten:

(8-8) $$(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Wir wollen nun (willkürlich) folgende Entscheidungsregel festlegen: Wenn unser Stichprobenergebnis $\bar{x}_1 - \bar{x}_2$ einen Wert annimmt, der um mehr als $\pm$ 1,96 Standardabweichungen von dem laut Hypothese erwarteten Wert $(\mu_1 - \mu_2) = 0$ entfernt liegt, dann wollen wir annehmen, daß diese Hypothese falsch ist. Wenn das Stichprobenergebnis jedoch näher zum Erwartungswert liegt, wollen wir sie beibehalten.

Spielen wir zunächst den besonders einfachen (aber ziemlich irrelevanten) Fall durch, in dem uns die beiden Standardabweichungen, σ_1 und σ_2, in der Population bekannt sind (so daß wir den Standardfehler der Mittelwertdifferenz nicht schätzen müssen). Wie wir anhand unseres vollständigen Datensatzes leicht feststellen können, ist

(8-9) $$\sigma_1 = 6{,}891$$

$$\sigma_2 = 6{,}098$$

In den beiden Stichproben erhalten wir folgende Mittelwerte für die Mandatsdauer:

(8-10) $\bar{x}_1 = 6,420$

$\bar{x}_2 = 7,540$

Somit ergibt sich in den Stichproben eine Mittelwertdifferenz von 1,12 Jahren, die die nicht-adligen Abgeordneten durchschnittlich länger im Reichstag verbracht haben als die adligen. Bevor wir daraus irgendwelche Schlüsse hinsichtlich der Haltbarkeit unserer Hypothese ziehen, müssen wir gemäß der zuvor vereinbarten Entscheidungsregel feststellen, um wieviel Standardabweichungen diese realisierte Differenz vom hypothetischen Erwartungswert $\mu_1 - \mu_2 = 0$, entfernt liegt. Unter Anwendung von (8-7) ergibt sich folgender z-Wert:

$$(8\text{-}11) \qquad z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{(6,42 - 7,54) - 0}{\sqrt{47,5/50 + 37,2/50}}$$

$$= \frac{-1,12}{\sqrt{0,95 + 0,744}} = \frac{-1,12}{1,30}$$

$$= -0,86$$

Der Standardfehler $\sigma_{\bar{x}_1 - \bar{x}_2}$ beträgt 1,30 Jahre. Die Stichprobendifferenz liegt somit um $|0,86|$ Standardabweichungen vom erwarteten Wert Null entfernt. Das entspricht laut z-Tabelle (Anhang A, Tabelle 1a oder b) einem $\alpha / 2$ - Wert von $|0,195|$. Gemäß unserer obigen Entscheidungsregel liegt er damit zu nahe am Erwartungswert, als daß wir die Hypothese (8-6) zurückweisen könnten. Wenn die Hypothese richtig ist, können wir erwarten, daß in ca. 39 % der Stichproben des angegebenen Umfanges Mittelwertdifferenzen $(\bar{x}_1 - \bar{x}_2) \geqq |1,12|$ realisiert werden.

Auch hier läßt sich wiederum ein (95%-)Vertrauensintervall berechnen:

$$(8\text{-}12) \qquad (\bar{x}_1 - \bar{x}_2) \pm z_{0,025}\sigma_{\bar{x}(1)-\bar{x}(2)} = -1,12 \pm 1,96 \cdot 1,30$$

$$= -1,12 \pm 2,548$$

Das Vertrauensintervall mit einer Irrtumswahrscheinlichkeit von $\alpha = 0,05$

schließt somit den Wert $\mu_1 - \mu_2 = 0$ ein. Erst wenn wir eine Irrtumswahrscheinlichkeit von $\alpha \geqq 0{,}39$ akzeptieren wollten, erhielten wir ein Konfindenzintervall, das eine Mittelwertdifferenz von Null nicht mehr einschließen würde. Andererseits kann man sich auch überlegen, wie groß die Stichprobenumfänge sein müßten, damit eine in den Stichproben beobachtete Mittelwertdifferenz $\bar{x}_1 - \bar{x}_2 = |1{,}12|$ bei unveränderter Entscheidungsregel zur Ablehnung der Hypothese (8-6) führen würde. Damit das Konfidenzintervall bei einem akzeptalen Fehlerrisiko von $\alpha = 0{,}05$ den Wert 0 nicht mehr einschließt, muß

$$(8\text{-}12') \qquad z_{0,025} \cdot \sigma_{\bar{x}(1)-\bar{x}(2)} \leq \bar{x}_1 - \bar{x}_2 \qquad , \sigma^2_{\bar{x}(1)-\bar{x}(2)} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

sein. Daraus folgt bei $n_1 = n_2 = n$

$$(8\text{-}12'') \qquad \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \leq \frac{1{,}12}{z_{0,025}} = 0{,}57$$

$$(\sigma_1^2 + \sigma_2^2)/n \leq 0{,}57^2 = 0{,}325$$

$$n \geq \frac{47{,}5+37{,}2}{0{,}325} = 261{,}34$$

Jede der beiden Teilstichproben mußte also mindestens 262 Fälle unfassen, um bei der betrachteten Mittelwertdifferenz von $\bar{x}_1 - \bar{x}_2 = 1{,}12$ die Hypothese (8-6) mit einem Fehlerrisiko $\alpha \leqq 0{,}05$ zurückweisen zu können.

Wenden wir uns nun dem realistischeren Fall zu, daß die Populationsvarianzen nicht nur ungleich, sondern auch unbekannt sind und somit über die modifizierten Stichprobenvarianzen geschätzt werden müssen. Sofern die Stichprobenumfänge »ausreichend groß« sind (was wir für unser Beispiel annehmen) und nicht stark differieren, ergeben sich keine Probleme. Die Prüfgröße

$$(8\text{-}13) \qquad t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}} = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}}$$

ist unter diesen Bedingungen, $(n_1 + n_2) \geqq 50$, annähernd normalverteilt[2]. Wir ermitteln in unserem Beispiel die (modifizierten) Stichprobenvarianzen

$$(8\text{-}14) \qquad \hat{\sigma}_1^2 = 38,3$$

$$\hat{\sigma}_2^2 = 35,1$$

Folglich ist

$$(8\text{-}15) \qquad t = \frac{-1,12}{\sqrt{\frac{38,3}{50}+\frac{35,1}{50}}} = \frac{-1,12}{-1,22}$$

$$= 0,918$$

Wenn wir an unserer Entscheidungsregel festhalten, führt auch dieses Ergebnis nicht zu einer Ablehnung der Hypothese. Der dort als Kriterium festgelegte z-Wert von | 1,96 |, der überschritten werden müßte, bleibt gültig, da wir unterstellt haben, daß bei den angegebenen Stichprobenumfängen die t-Verteilung der Normalverteilung entspricht.

Etwas komplizierter wird die Sachlage, wenn bei ungleichen Populationsvarianzen die Stichprobenumfänge auch noch klein oder/und sehr unterschiedlich groß sind. Man kann dann zwar ebenfalls die Prüfgröße (8-15) berechnen; für den t-Test ist aber noch eine Korrektur der Freiheitsgrade nach folgender Formel (»Welch-Test«) vorzunehmen (siehe Schlittgen 1987, S. 329; Hays 1973, S. 409f.):

---

[2] In Gleichung (8-13) gehen wir von der Definition $\hat{\sigma}^2 = 1/(n-1)\,\Sigma\,(x_i-\bar{x})^2$ aus. Es besteht aber die Identität: $\hat{\sigma}^2/n = s^2/n-1$ mit $s^2 = 1/n\,\Sigma\,(x_i-\bar{x})^2$, so daß man in der Literatur beide Schreibweisen findet.

$$(8\text{-}16) \qquad df = \frac{(1-R)^2}{R_2/(n_1-1) + 1/(n_2-1)}$$

$$R \text{ steht für } \frac{\hat{\sigma}_x^2/n_1}{\hat{\sigma}_y^2/n_2}$$

Sehen wir uns zum Schluß noch an, welche Beziehung nun tatsächlich in der Grundgesamtheit gegeben ist. Wir erhalten $\mu_1 = 7{,}481$ und $\mu_2 = 7{,}035$. Anders als in der Stichprobe sind in der Grundgesamtheit die adligen Abgeordneten im Durchschnitt längere Zeit im Reichstag als die nicht-adligen Abgeordneten. Die Differenz beträgt aber nicht einmal ein halbes Jahr. Sie liegt mit +0,446 Jahren noch innerhalb des 95 %-Vertrauensintervalls, das wir gemäß den Gleichungen (8-11) bis (8-15) aus den Stichprobendaten schätzen können. Das Beispiel zeigt, daß Punktschätzungen auf der Basis von Stichprobendaten mit großer Vorsicht zu genießen sind, wenn die Stichprobenumfänge klein sind. Auf jeden Fall sollten Vertrauensintervalle berechnet werden.

8.3 Wünschenswerte Eigenschaften von Schätzfunktionen

In vorangegangenen Abschnitten haben wir ziemlich naiv, allein auf der Basis einer formalen Analogie, das Stichprobenmittel $\bar{x}$ als »Schätzer« für das arithmetische Mittel μ der Population und die (leicht modifizierte) Stichprobenstreuung s^2 als Schätzer für die Varianz σ^2 der Population benutzt. Da diese »Punktschätzer« von Stichprobe zu Stichprobe mehr oder weniger stark um ihre »Zielgröße« (den entsprechenden Parameter der Population oder, allgemeiner, einer theoretischen Verteilung) streuen, schien es sinnvoll, Vertrauensintervalle zu berechnen. In diesem Abschnitt wollen wir formale Gütekriterien erörtern, die die Auswahl der jeweiligen Schätzer besser begründen können als eine formale Analogie alleine. Die Notwendigkeit solcher Kriterien wird schon erkennbar, wenn wir uns fragen, warum wir bei symmetrischen Verteilungen das arithmetische Mittel $\bar{x}$ und nicht den Median $\tilde{x}$ als Schätzer für μ verwenden. Schließlich sind beide in symmetrischen Verteilungen identisch.

Wir wollen zunächst vereinbaren, beliebige Schätzer (sei es $\bar{x}$, s^2, ein Korrelationskoeffizient oder eine andere Kennzahl) mit dem Kürzel $\hat{\theta}$ zu bezeichnen und die zu schätzenden Parameter mit θ. Da es sich bei einem Schätzer stets um eine Stichprobenfunktion handelt, spricht man statt von

»Schätzern« auch von »Schätzfunktionen«. Man sagt: der realisierte Wert $\hat{\theta}(x_1, \ldots, x_n)$ einer Stichprobenfunktion $\hat{\theta}(X_1, \ldots, X_n)$ dient als Schätzwert für einen theoretischen Parameter θ.

Das wichtigste Konzept zur Beurteilung einer Schätzfunktion ist der mittlere quadratische Fehler (MQF oder MSE = Mean Square Error). Der MQF gibt an, mit welchen quadrierten Abständen zwischen dem zu schätzenden Parameter und seinem Schätzwert im Mittel (bei sehr vielen Stichprobenziehungen) zu rechnen ist:

$$(8\text{-}17) \qquad MQF(\hat{\theta},\theta) = E[(\hat{\theta}(x_1,\ldots,x_n)-\theta]^2$$

Es wird derjenige Schätzer $\hat{\theta}$ vorgezogen, der »auf lange Sicht« erwarten läßt, daß er näher an der Zielgröße θ liegt als irgendein »Konkurrent«. Obwohl die Zielgröße unbekannt ist, können wir mit Hilfe unseres Wissens über die (theoretische) Verteilung der Stichprobenfunktion $\hat{\theta}$ den MQF bestimmen. Er setzt sich aus zwei Komponenten zusammen, dem Quadrat der »Verschiebung« von $\hat{\theta}$ gegenüber θ und der Varianz von $\hat{\theta}$:

$$(8\text{-}18) \qquad MQF(\hat{\theta},\theta) = [E(\hat{\theta})-\theta]^2 + V(\hat{\theta})$$

Diese Beziehung erhält man, wenn man (8-17) umformt zu

$$(8\text{-}19) \qquad MQF(\hat{\theta},\theta) = E(\hat{\theta}-\theta)^2 = E[\hat{\theta}-E(\hat{\theta}) + E(\hat{\theta})-\theta]^2$$

und das Binom auf der rechten Seite ausmultipliziert.

$$(8\text{-}19') \qquad MQF = E[\hat{\theta}-E(\hat{\theta})]^2 + E[E(\hat{\theta})-\theta]^2 + 2E[(\hat{\theta}-E(\hat{\theta})(E(\hat{\theta})-\theta)]$$

$$= V(\hat{\theta}) + [E(\hat{\theta})-\theta]^2 + 0$$

Da $E(\hat{\theta})$ wie auch θ eine Konstante ist, die Differenz zweier Konstanten wiederum eine Konstante ergibt und der Erwartungswert einer Konstante gleich der Konstante ist, können wir für $E[E(\hat{\theta})-\theta]^2$ auch $[E(\hat{\theta}-\theta)]^2$ schreiben. Der letzte Ausdruck wird Null, da $E[\hat{\theta}-E(\hat{\theta})]=0$.

Die Verschiebung wird auch als »Bias« oder »Verzerrung« bezeichnet. Sie gibt an, in welchem Maße der Schätzer bei »sehr vielen« Stichprobenziehungen »im Mittel« von dem »wahren« Parameter abweicht. Das erwartbare Fehlerquadrat ist aber auch davon abhänig, wie stark der Schätzer bei wiederholter Stichprobenziehung um diesen Parameter streut. Es kann durchaus sein, daß zwar keinerlei »Bias« vorliegt, aber dennoch aufgrund der großen Streubreite des Schätzers ein erheblicher MQF zu erwarten ist.

Die Abbildung 8.2 veranschaulicht das Zusammenwirken dieser beiden Komponenten. Der Schätzer $\hat{\theta}_1$ ist nicht »verschoben« (sein Erwartungswert stimmt mit der Zielgröße θ überein), streut aber relativ breit. $\hat{\theta}_2$ weist eine kleine Verschiebung auf, ist aber relativ stark um den (eigenen) Er-

wartungswert konzentriert. $\hat{\theta}_3$ schließlich ist zwar relativ stark konzentriert, weist aber eine sehr große Verschiebung auf. Eine Schätzfunktion, die keinen Bias aufweist: $E(\hat{\theta}) - \theta = 0$, heißt **erwartungstreu** (unverzerrt). Diejenige Schätzfunktion, die eine minimale Varianz aufweist, heißt **effizient**. Die **relative Effizienz** eines Schätzers $\hat{\theta}_1$ im Vergleich zu einem anderen Schätzer, $\hat{\theta}_2$ ist

$$(8\text{-}20) \qquad \frac{E(\hat{\theta}_2 - \theta)^2}{E(\hat{\theta}_1 - \theta)^2}$$

Da $E(\hat{\theta} - \theta)^2 = V(\hat{\theta})$, kann man auch sagen: Die relative Effizienz ist das Verhältnis der quadrierten Standardfehler der beiden Schätzer.

Erwartungstreue und Effizienz potentieller Schätzer lassen sich aus ihren Verteilungs- bzw. Dichtefunktionen bestimmen. Für die Stichprobenfuktionen $\hat{\theta}_1 = \bar{x}$ und $\hat{\theta}_2 = \tilde{x}$ als Schätzgrößen für das arithmetische Mittel μ sieht das bei vorausgesetzter symmetrischer Verteilung der Variablen X wie folgt aus:

Der Zentrale Grenzwertsatz besagt, daß $\bar{x}$ asymptotisch normalverteilt ist mit $E(X) = \mu$ und $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$. Auch der Median $\tilde{x}$ ist bei stetigen, symmetrischen, unimodalen Verteilungen im allgemeinen asymptotisch normalverteilt um den Erwartungswert μ mit der Varianz $\Sigma\ \sigma^2/2n = 1{,}57(\sigma^2/n)$. Der Median, obwohl erwartungstreu, hat also eine etwa eineinhalb mal so große Varianz wie das arithmetische Mittel. Die relative Effizienz von $\bar{x}$ im Verhältnis zu $\tilde{x}$ ist somit 157 %:

$$(8\text{-}21) \qquad \frac{\frac{\pi\sigma^2}{2n}}{\frac{\sigma^2}{n}} = \frac{\pi\sigma^2 n}{2n\sigma^2} = \frac{\pi}{2} = 1{,}57$$

Folglich ist das arithmetische Mittel dem Median als Schätzer für μ bei symmetrischen Verteilungen vorzuziehen: Im Schnitt ist $\bar{x}$ näher am Zielwert »dran«; er liefert somit ein präziseres (engeres) Vertrauensintervall für μ bei gleicher Irrtumswahrscheinlichkeit. Das ändert sich aber, wenn wir es mit einer Verteilung zu tun haben, bei der extreme Werte mit großer Wahrscheinlichkeit vorkommen. Extreme Werte beeinflussen das arithmetische Mittel sehr stark. Folglich ist in diesem Falle $\tilde{x}$ ein effizienterer Schätzer als $\bar{x}$.

Eine weitere Eigenschaft, die man sich bei Schätzern wünscht, ist die **Konsistenz**. Man nennt einen Schätzer konsistent, wenn bei ihm mit zunehmendem Stichprobenumfang sowohl der Bias verschwindet als auch die Varianz gegen Null strebt, wenn also

$$(8\text{-}22) \qquad \lim_{n \to \infty} MQF(\hat{\theta}, \theta) = 0$$

Die Werte einer konsistenten Schätzfunktion konzentrieren sich also mit zunehmendem Stichprobenumfang immer enger um den wahren Wert des zu schätzenden Parameters 0. Mit der Konsistenz wird dem Statistiker versichert, daß er bei größerem Aufwand (größerer Stichprobe) auch bessere Ergebnisse erwarten kann. Die Konsistenz wird als minimales Gütekriterium aufgefaßt. Wenn ein Schätzer nicht einmal konsistent ist, ist er in der Regel nicht zu empfehlen.

Zu den wünschenswerten Eigenschaften eines Schätzers zählt auch seine **Robustheit**. Ein Schätzer wird als robust bezeichnet, wenn er nicht empfindlich auf Ausreißerwerte oder die Verletzung von Modellannahmen (z. B. der Normalverteilungsannahme) reagiert. Wir können hier aber nicht die Methoden erläutern, mit denen man die Robustheit eines Schätzers ermitteln kann (siehe Schlittgen 1987, 264 ff.).

Die wichtigsten Verfahren zur Konstruktion von Schätzfunktionen sind die »Methode der kleinsten Quadrate« und die »Maximum-Likelihood-Methode«; beide werden wir im Zusammenhang mit der Regressionsanalyse kennenlernen.

8.4 Schätzen von Konfidenzintervallen

Alle Schätzer, auch die erwartungstreuen und effizienten, schwanken von Stichprobe zu Stichprobe. Deshalb ist es sinnvoll, nicht nur eine Punktschätzung vorzunehmen, sondern Vertrauensintervalle mit einer vorgegebenen Irrtumswahrscheinlichkeit zu bestimmen, wie wir das in Abschnitt 8.1 für das arithmetische Mittel gezeigt haben. Das Konfidenzintervall wird als ein Paar von Zufallsvariablen betrachtet. $\bar{x} - z_{\alpha/2} \cdot \sigma_{\bar{x}}$ als das untere Ende und $\bar{x} + z_{\alpha/2} \cdot \sigma_{\bar{x}}$ als das obere Ende des Vertrauensintervalls für μ sind Stichprobenfunktionen, d. h., sie hängen, da σ, n und α gegeben sind, nur von den Stichprobenvariablen $X_1, \ldots, X_n$ und deren Realisierungen ab.

So, wie wir das Vertrauensintervall für μ bestimmt haben, werden im Prinzip auch die Vertrauensintervalle anderer Parameter ermittelt. Man braucht zunächst eine geeignete Schätzfunktion, deren Verteilung und Standardfehler. Es muß sichergestellt sein, daß die Voraussetzungen erfüllt sind, unter denen das Verteilungsmodell gilt (zum Beispiel: liegt eine ein-

fache Zufallsstichprobe vor oder nicht?). In der Regel bezieht sich das Verteilungsmodell auf eine transformierte Stichprobenfunktion. So haben wir $\bar{x}$ zu z oder t, s^2 zu χ^2 transformiert. Mit der Transformation wird angestrebt, die Verteilungsfunktion des Schätzers $\hat{\theta}$ unabhängig von dem gegebenen (aber unbekannten) Parameter θ zu machen. Wir wollen im folgenden die transformierte Stichprobenfunktion mit T abkürzen.

Im nächsten Schritt muß eine Irrtumswahrscheinlichkeit α für das Vertrauensintervall gewählt werden. Dadurch sind die T-Werte $T_{\alpha/2}$ und $T_{1-\alpha/2}$ bestimmt, die laut Verteilungsmodell die Grenzen desjenigen Intervalls markieren, in das bei wiederholten Ziehungen $100(1-\alpha)$ Prozent Realisierungen der (transformierten) Stichprobenfunktion fallen. (Wie bereits erwähnt ist bei symmetrischen Verteilungen $T_{\alpha/2} = -T_{1-\alpha/2}$) Das läßt sich in Form einer Ungleichung ausdrücken:

$$(8\text{-}23) \qquad P(T_{\alpha/2} \leq T(\hat{\theta}) \leq T_{1-\alpha/2}) = 1 - \alpha$$

$T(\hat{\theta})$ soll dabei den transformierten realisierten Schätzwert der Stichprobenfunktion $\hat{\theta}(X_1, \ldots, X_n)$ darstellen, z. B.

$$(8\text{-}24) \qquad \hat{\theta} = \bar{x} \rightarrow T(\hat{\theta}) = \frac{\bar{x}-\mu}{\sigma_x/\sqrt{n}} = z_x$$

Die Ungleichung in der Klammer von (8-23) läßt sich so umformen, daß in ihrem Zentrum der interessierende Parameter θ isoliert wird. Wir wollen das im folgenden noch einmal für das (transformierte) arithmetische Mittel (normalverteilt) und die (transformierte) Varianz (χ^2-verteilt) zeigen:

$$(8\text{-}25) \qquad P(z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

$$P(-1{,}96 \leq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \leq 1{,}96) = 1 - \alpha$$

Multiplikation mit dem Nennerausdruck:

$$P(-1{,}96 \cdot \sigma_{\bar{x}} \leq \bar{x} - \mu \leq 1{,}96 \cdot \sigma_{\bar{x}}) = 1 - \alpha$$

Subtraktion von $\bar{x}$:

$$P(-1{,}96 \cdot \sigma_{\bar{x}} - \bar{x} \leq -\mu \leq 1{,}96 \cdot \sigma_{\bar{x}} - \bar{x}) = 1 - \alpha$$

Multiplikation mit (-1), Umkehrung der Vorzeichen:

$$P(\bar{x} + 1{,}96 \cdot \sigma_{\bar{x}} \geq \mu \geq \bar{x} - 1{,}96 \cdot \sigma_{\bar{x}}) = 1 - \alpha$$

$$P(\bar{x} - 1{,}96 \cdot \sigma_{\bar{x}} \leq \mu \leq \bar{x} + 1{,}96 \cdot \sigma_{\bar{x}}) = 1 - \alpha$$

Damit haben wir das Ergebnis von (8-2) wiederholt, um es leichter mit der folgenden Entwicklung bezüglich der Varianz vergleichen zu können. Dabei muß beachtet werden, daß die Chi-Quadrat-Werte im Unterschied zu den z-Werten von den jeweils vorliegenden Freiheitsgraden n-1 abhängen.

$$(8\text{-}26) \qquad P\left(\chi_{df;\alpha/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{df;1-\alpha/2}^2\right) = 1 - \alpha$$

Auch diese Ungleichung muß wieder so umgeformt werden, daß der interessierende Parameter, σ^2, isoliert wird. Das ist jetzt ein bißchen komplizierter als vorher, weil er im Nenner steht. Im ersten Schritt bilden wir deshalb die Reziprokwerte, was erfordert, die Ungleichheitszeichen umzukehren:

$$(8\text{-}27) \qquad P\left(\frac{1}{\chi^2_{df;\alpha/2}} \geq \frac{\sigma^2}{(n-1)s^2} \geq \chi^2_{df;1-\alpha/2}\right) = 1 - \alpha$$

Nun multiplizieren wir alle Ausdrücke mit $(n-1)s^2$:

$$(8\text{-}28) \qquad P\left(\frac{(n-1)s^2}{\chi^2_{df;\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)s^2}{\chi^2_{df;1-\alpha/2}}\right) = 1 - \alpha$$

Um wieder eine Ungleichung zu erhalten, in der die kleinste Größe ganz links steht, schreiben wir

$$(8\text{-}29) \qquad P\left(\frac{(n-1)s^2}{\chi^2_{df;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{df;\alpha/2}}\right) = 1 - \alpha$$

Wenn wir für die Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ wählen und 30 Freiheitsgrade (df) zur Verfügung haben, erhalten wir (siehe die Tabellen 3a und 3b im Anhang A)

$$(8\text{-}30) \qquad P\left(\frac{30s^2}{46{,}98} \leq \sigma^2 \leq \frac{30s^2}{16{,}79}\right) = 0{,}95$$

Beim Ablesen der χ^2-Werte aus den Tabellen stoßen wir auf das gleiche Problem wie bei der tabellierten Standardnormalverteilung: In den Tabellen der meisten Lehrbücher sind die Dichtefunktionen von »rechts nach links« integriert, die χ^2-Werte nehmen also mit kleinerem Wahrscheinlichkeitsgrad zu, während wir bei unserer Notation die »normale« Integration von links nach rechts vorausgesetzt haben. Deshalb sind im Anhang zwei Chi-Quadrat-Tabellen wiedergegeben; Tabelle 3a entspricht unserer Notation, Tabelle 3b der anderen, die sich direkt an der Praxis der »Signifikanztests« (siehe nächsten Abschnitt) orientiert.

8.5 Zur Logik des Testens von Hypothesen

Eine Hypothese, die statistisch getestet werden soll, ist nach einer Definition von Blalock (1960, S.91) »a statement about a future event, or an event the outcome of which is unknown at the time of the prediction, set forth in such a way that it can be rejected«. Nicht alle Theorien sind direkt testbar. In den empirischen Wissenschaften sollten sie aber zu Folgerungen führen, die ihrerseits (direkt) testbar sind und auf diese Weise die Theorie bestätigen oder widerlegen können. Laut Blalock (ebd.) erfordert der Hypothesentest folgende Schritte:

1. Alle möglichen Ergebnisse eines Experiments oder einer Beobachtung werden antizipiert, bevor der Test durchgeführt wird.
2. Schon vor dem Test wird festgelegt, mit welchen Operationen oder Prozeduren festzustellen ist, welche Ergebnisse tatsächlich aufgetreten sind.
3. Es wird im voraus entschieden, aufgrung welcher Ergebnisse, falls sie auftreten, die Hypothese zurückgewiesen wird und aufgrund welcher Ergebnisse sie nicht zurückgewiesen wird. Die Zurückweisung muß ein mögliches Resultat sein.
4. Das Experiment wird durchgeführt bzw. das Ergebnis wird beobachtet, die Ergebnisse werden festgehalten und es wird eine Entscheidung darüber gefällt, ob die Hypothese zurückgewiesen ist oder nicht.

Diese Regeln sollen vor allem sicherstellen, daß die Kriterien, nach denen eine Hypothese angenommen oder abgelehnt wird, offengelegt und nicht nachträglich verändert werden, wenn das Testergebnis bereits vorliegt.

8.5.1 Formulierung von Forschungs- und Nullhypothese

Viele der theoretisch interessierenden (testbaren) Hypothesen lassen sich als Behauptungen über Zusammenhänge zwischen Variablen operationalisieren, deren Form und Stärke mit Hilfe bestimmter »Koeffizienten« (statistischer Maßzahlen) dargestellt werden können. So haben wir in Abschnitt 8.2 den (Nicht-)Zusammenhang zwischen den Variablen »Mandatsdauer« und »Zugehörigkeit/Nichtzugehörigkeit zum Adel« mit Hilfe einer Mittelwertdifferenz dargestellt.

Häufig ist der Forscher aber nicht in der Lage, eine Hypothese so weit zu präzisieren, daß sie in der oben erläuterten Weise testbar wäre. Er mag z. B. ziemlich sicher sein, daß zwischen zwei Variablen X und Y ein Zusammenhang besteht, sieht aber keine Möglichkeit, die Stärke des vermuteten Zusammenhangs z. B. mit $r = 0{,}35$ oder $d\% = 27$ genau anzugeben. Statistisch, anhand von Stichprobendaten, läßt sich eine Hypothese im Sinne des hier zu erläuternden Signifikanztests aber nur dann testen, wenn sie als eine präzise Behauptung über den vermuteten Wert einer statistischen Kennzahl operationalisiert werden kann. Man spricht dann auch von »spezifischen« im Unterschied zu »unspezifischen « Hypothesen. Was also ist zu tun, wenn wir nicht in der Lage sind, eine derart präzise Forschungshypothese zu begründen? Die Antwort heißt: indirektes Testen. Machen wir uns am Beispiel der durchschnittlichen Mandatsdauer adliger und nicht-adliger Reichstagsabgeordneter deutlich, was das heißt.

Im Abschnitt 8.2 hatten wir die Hypothese getestet, daß kein Zusammenhang zwischen der durchschnittlichen Mandatsdauer der Abgeordneten und ihrer Zugehörigkeit oder Nichtzugehörigkeit zur Adelsschicht bestehe. Die »erwartete« Mittelwertdifferenz war also gleich Null. Gehen wir nun einmal von der gegenteiligen Forschungshypothese aus, daß zwischen diesen beiden Variablen ein Zusammenhang besteht. Auch diese Hypothese kann wiederum mit Hilfe einer Aussage über die Mittelwertdifferenz operationalisiert werden:

$$(8\text{-}31) \qquad H_1: (\mu_1 - \mu_2) \neq 0$$

(Der Index 1 bezieht sich wieder auf die Adligen, der Index 2 auf die Nicht-Adligen.)

Wenn wir uns daran machen, **diese** Hypothese zu testen, stoßen wir sofort auf ein schwerwiegendes Problem: Welchen Erwartungswert $E(\bar{x}_1 - \bar{x}_2)$ sollen wir ansetzen? H_1 nennt gar keinen spezifischen Erwartungswert. Ohne Angabe eines Erwartungswertes für die betreffende statistische Kennzahl (hier: Mittelwertdifferenz) ist jedoch der übliche Signifikanztest, auf den wir uns hier ausschließlich beziehen wollen, nicht möglich[3]. Einen Ausweg aus diesem Problem bietet die logische Negation der Hypothese H_1:

[3] Aus der Perspektive der statistischen Entscheidungstheorie lassen sich

(8-32) $H_0: (\mu_1 - \mu_2) = 0$

Das ist genau diejenige Hypothese, die wir schon in Abschn. 8.2 getestet haben. Sie beinhaltet die präzise Angabe des Erwartungswertes $E(\bar{x}_1 - \bar{x}_2) = 0$. Wenn wir in der Lage sind, diese Hypothese zurückzuweisen, impliziert dies eine Bestätigung der »eigentlichen« Forschungshypothese H_1. Wenn H_0 falsch ist, muß die logische Negation H_1 richtig sein - und umgekehrt: Wenn die Hypothese, es besteht keine Mittelwertdifferenz richtig ist, muß die Hypothese, es bestehe eine Mittelwertdifferenz, falsch sein. Über den Test der präzisen Hypothese H_0 testen wir also indirekt die unpräzise Hypothese H_1. Eine Unsicherheit ist jedoch nicht zu umgehen. Wenn wir eine Hypothese anhand von Stichprobendaten, also mit Hilfe von Schätzfunktionen testen, sind unsere Entscheidungen über »richtig« oder »falsch« unweigerlich mit einem Fehlerrisiko behaftet. Art und Ausmaß des Fehlerrisikos werden wir sogleich besprechen, wollen aber zuvor noch eine Bemerkung zur Terminologie anbringen:

Man nennt diejenige Test-Hypothese, die einen präzisen Wert θ_0 für den theoretisch interessierenden Parameter θ angibt, eine **Nullhypothese**. Ihr (unpräzises) »Gegenstück«, ihre logische Negation, nennt man **Alternativhypothese** oder auch »Arbeitshypothese«, »Forschungshypothese«, »Gegenhypothese«. Den Vorgang, mit dem man auf Stichprobenbasis über Beibehaltung oder Ablehnung der Nullhypothese entscheidet, nennt man einen **Signifikanztest**. In einigen Lehrbüchern werden auch »substantielle« Definitionen von Alternativ- und Nullhypothese vorgeschlagen. Es wird z. B. gesagt, die Alternativhypothese sei diejenige, an deren Beibehaltung der Forscher »eigentlich« interessiert sei, weil sie über den bisher erreichten Wissensstand hinausführe. Dies entspricht zwar in der Tat häufig der Intention des Forschers, muß es aber nicht. Gelegentlich ist er auch an der Beibehaltung einer Hypothese interessiert, mit der er einen präzisen Erwartungswert formulieren und die er deshalb direkt testen kann. Deshalb ist es sinnvoller die Nullhypothese rein formal im Hinblick auf ihre Funktion im Testverfahren zu definieren: Sie ist diejenige Hypothese, die wegen ihrer Präzision direkt getestet werden kann. Unter der Alternative H_1 werden dagegen alle Parameterwerte zusammengefaßt, die ebenfalls möglich sind, die aber nicht zu H_0 gehören.

An dieser Stelle muß ein weiteres Problem erwähnt werden: Häufig gelingt es dem Forscher, seine Arbeitshypothese etwas präziser zu formulieren, als wir das eben in unserem Beispiel getan haben, ohne sich auf einen bestimmten Erwartungswert festlegen zu können. Statt der Hypothese

---

neben Punktwerten auch beliebige Wertebereiche testen. Auf diese allgemeine Konzeption statistischer Tests wollen wir hier nicht eingehen. Vgl. dazu Mood/Graybill/Boes (1974, Kap. 9).

(8-33) $H_1: \mu_1 \neq \mu_2$

läßt sich in unserem Beispiel die Hypothese

(8-34) $H_1^*: \mu_1 > \mu_2$

vertreten. Man kann z. B. darauf verweisen, daß viele adlige Abgeordnete ostelbischen Familien mit Großgrundbesitz entstammen und daß die mit dem Großgrundbesitz verbundenen Patronagefunktionen und die höhere »Abkömmlichkeit« eine längere Mandatsdauer begünstigen. Auch die Hypothese H_1^* ist nicht direkt testbar. Jetzt ist allerdings auch die logische Negation

(8-35) $H_0^*: \mu_1 \leq \mu_2$

unpräzise und scheinbar nicht zu testen. Wie man sich aus diesem Problem befreien kann, zeigen wir in Abschnitt 8.5.4.

8.5.2 Fehlertypen und Signifikanzniveau

Im vorigen Abschnitt haben wir erläutert, wie Nullhypothese und Alternativhypothese so formuliert werden, daß sie gemeinsam alle theoretisch möglichen Werte des interessierenden Parameters (z. B. Mittelwertdifferenzen) umfassen. Über die Beibehaltung oder Ablehnung der Hypothese entscheidet man auf Grund eines Stichprobenergebnisses nach zuvor festgelegten Kriterien. Bei dieser Entscheidung können Fehler auftreten:

(1) Man entscheidet sich gegen die Nullhypothese (und damit für die Alternativhypothese), obwohl sie richtig ist. Diese Fehlentscheidung bezeichnet man als »Fehler erster Art« oder »Alpha-Fehler«.
(2) Man entscheidet sich für die Nullhypothese (und damit gegen die Alternativhypothese), obwohl sie falsch ist. Diese Fehlentscheidung wird als »Fehler zweiter Art« oder »Beta-Fehler« bezeichnet.

Richtige und falsche Entscheidungsmöglichkeiten sind in der folgenden Übersicht zusammengefaßt:

		In der Population gilt die:	
		H_0	H_1
Entscheidung auf Grund der Stichprobe zugunsten der:	H_0	richtige Entscheidung	β-Fehler
	H_1	α-Fehler	richtige Entscheidung

Quelle: Bortz 1979, S. 142

Da wir den Test durchführen, weil der interessierende Populationsparameter unbekannt ist, können wir im Einzelfall nie wissen, ob wir eine richtige oder falsche Entscheidung getroffen haben. Aber immerhin können wir definitive Aussagen über die bedingten Wahrscheinlichkeiten von Fehlentscheidungen machen:

Wir haben in unserem Beispiel zur durchschnittlichen Mandatsdauer der adligen und der nicht-adligen Reichstagsabgeordneten bereits gezeigt, wie man Entscheidungsregeln formuliert. Man geht vorläufig von der Richtigkeit der Nullhypothese aus und legt fest, wie weit entfernt von dem hypothetisch erwarteten Wert das Stichprobenergebnis liegen muß, wenn es zur Ablehnung der Nullhypothese führen soll. Man unterteilt also den Bereich möglicher Ergebnisse (realisierbarer Werte) in einen Ablehnungs- (oder »Verwerfungs«-) und einen Annahmebereich nach dem Muster von Abb. 8.4 (bezogen auf unser Beispiel).

Denjenigen Punkt auf der (transformierten) Parameterskala, an dem die beiden Bereiche aneinandergrenzen, bezeichnet man als »kritischen Wert«. Bei der Standardnormalverteilung, einer Hypothese vom Typ (8-32) und einem vorgegebenen Alpha-Wert von, beispielsweise, $\alpha = 0{,}05$ sind $z_1 = -1{,}96$ und $z_2 = +1{,}96$ die beiden kritischen Werte. Wenn das Stichprobenergebnis in den Ablehnungsbereich fällt, wird die Nullhypothese zurückgewiesen. Diese Entscheidung kann aber falsch sein. Denn auch wenn die Nullhypothese richtig ist, werden in $100 \cdot \alpha$ Prozent der Fälle (bei wiederholter Stichprobenziehung) Stichprobenergebnisse erzielt, die soweit vom Erwartungswert entfernt liegen, daß sie in den Ablehnungsbereich fallen, der durch den Alpha-Wert vor Durchführung des Tests definiert worden ist. Man bezeichnet deshalb die Größe Alpha als **Fehlerrisiko** oder **Irrtumswahrscheinlichkeit**. Dabei handelt es sich um eine **bedingte** Wahrscheinlichkeit für das Auftreten eines extremen Ereignisses. Die angenommene Bedingung dabei ist, daß die Nullhypothese zutrifft.

Mit dem Alpha-Fehler ist das sog. **Signifikanzniveau** bezeichnet. Wenn man z. B. sagt: Die Korrelation zwischen der Variable X und der Variable

Y sei auf dem 5-Prozent-Niveau »signifikant«, bedeutet dies, daß die Nullhypothese des Nichtzusammenhangs mit einem Fehlerrisiko von $\alpha \leqq 0{,}05$ zurückgewiesen werden konnte. Dies ist aber eine rein statistische Feststellung, die über die tatsächliche Stärke des Zusammenhangs und deren theoretische Relevanz nichts aussagt.

Gemäß unserer Entscheidungsregel wird die Nullhypothese nicht verworfen, wenn das Stichprobenergebnis in den durch Alpha festgelegten Annahmebereich fällt. Aber auch in diesem Falle kann eine Fehlentscheidung vorliegen. Denn in den Annahmebereich können Stichprobenergebnisse auch dann fallen, wenn die Nullhypothese falsch ist. Leider jedoch kann man die Wahrscheinlichkeit ß, diesen Fehler zu begehen, im konkreten Einzelfall nicht bestimmen, da sie von dem unbekannten Populationsparameter abhängt. Das kann man sich an der Abbildung 8.5 klarmachen. (Wir bezeichnen den Populationsparameter mit θ und das realisierte Stichprobenergebnis mit $\hat{\theta}$)

Wenn die Nullhypothese richtig ist, gibt es die Wahrscheinlichkeit $\alpha/2$ (senkrecht schraffierte Fläche), daß ein Wert $\hat{\theta}$ realisiert wird, der mindestens so groß ist wie in der Abbildung eingetragen. Wenn aber in der Population entgegen der Nullhypothese der Parameterwert θ_1 realisiert ist, gibt es einen durch das $ß_1$-Feld (einfach und doppelt schraffierte Flächen links von $\hat{\theta}$) markierte Wahrscheinlichkeit dafür, daß ein Stichprobenergebnis realisiert wird, das in den Annahmebereich der Nullhypothese fällt. Das heißt, es gibt eine Wahrscheinlichkeit $ß_1$, daß die Nullhypothese beibehalten wird, obwohl sie falsch ist. Wenn aber in der Population, aus der die Stichprobe gezogen wurde, nicht der Parameterwert θ_1, sondern θ_2 gegeben wäre, wäre $ß_2 < ß_1$ (die Größe von $ß_2$ ist durch die doppelt schraffierte Fläche dargestellt). Da wir nicht wissen, welcher Wert: θ_1, θ_2, oder ein anderes θ_k, in der Population gegeben ist, können wir keine präzise Angabe über den ß-Fehler machen. Wir können lediglich festhalten, daß unabhängig von dem wahren θ der ß-Fehler um so kleiner wird, je größer wir das Fehlerrisiko α wählen - und umgekehrt, daß der ß-Fehler um so größer wird, je kleiner α. Allerdings kann man nicht davon ausgehen, daß sich der ß-Fehler im gleichen Maße verringert, wie der α-Fehler erhöht wird.

Der Forscher steckt also in einem Dilemma. Der Sozialforscher umgeht es meistens in der Weise, daß er sich an die etablierten Konventionen und Faustregeln hält, die als Testkriterium ein Fehlerrisiko von $\alpha = 0{,}05$ oder 0,01 empfehlen. Diese Konventionen stützen sich auf den Tatbestand, daß die Nullhypothese sehr häufig diejenige Hypothese ist, die der Forscher gerne zurückweisen möchte. So ist er z. B. eher daran interessiert, eine (unpräzise) Zusammenhangshypothese zu bestätigen als die Nullhypothese eines Nichtzusammenhangs beizubehalten. Das tradierte Forschungsethos verlangt nun, daß er sich die Ablehnung der »ungeliebten« Nullhypothese

schwer macht, indem er ein kleines Alpha-Niveau festlegt. Folglich erhöhen viele Forscher ihr Alpha-Niveau auf $\alpha = 0{,}20$ oder noch höher, wenn die Nullhypothese diejenige ist, an deren Beibehaltung sie interessiert sind. Wenn die »gemochte« Hypothese als Alternativhypothese fungiert, wird sie um so strenger (indirekt) getestet, je kleiner das Alpha-Niveau für die Ablehnung der Nullhypothese festgelegt wird. Wenn die »gemochte« Hypothese so präzise ist, daß sie als Nullhypothese fungiert, wird sie um so strenger getestet, je größer das vorgegebene Alpha-Niveau ist.

Der Forscher bewegt sich bei der Wahl des Fehlerniveaus stets auf unsicherem Grund. Er muß letztlich aus »inhaltlichen« Erwägungen entscheiden, wie wichtig es ihm ist, auf keinen Fall eine falsche Forschungshypothese zu akzeptieren - oder: wie wichtig es ihm ist, überhaupt erst einmal eine Theorie auf die Beine zu stellen, auch wenn sie zunächst auf relativ schwachen Füßen steht, ihre Negation also bei konventionell gewähltem $\alpha = 0{,}05$ nicht zurückgewiesen werden kann.

Viele Lehrbücher empfehlen, in den Forschungsberichten nicht nur das vorgängig festgelegte Fehlerrisiko mitzuteilen, an der eine Hypothese gescheitert ist oder sich bewährt hat, sondern den im Test empirisch erreichten Wahrscheinlichkeitswert (»prob value«, »empirisches Signifikanzniveau«) anzugeben. Dieses empirische Signifikanzniveau α^* ist also diejenige Wahrscheinlichkeit, mit der eine Prüfgröße T (in unserem Beispiel war T die z-transformierte Mittelwertdifferenz) einen Wert $T = t_1$ annimmt, der laut Nullhypothese genauso oder noch weniger wahrscheinlich ist als der tatsächlich beobachtete Wert $T = t^*$. Im Unterschied zu dem a priori festgelegten Fehlerrisiko, das nicht überschritten werden soll, wenn man die Nullhypothese zurückweisen will, gibt das empirische Signifikanzniveau das Fehlerrisiko an, das man tatsächlich eingeht, wenn man die Nullhypothese aufgrund des beobachteten T-Wertes zurückweist[4]. Der Leser des Forschungsberichts kann dann selbst entscheiden, ob er ebenfalls bei diesem Signifikanzniveau die Nullhypothese zurückgewiesen oder beibehalten hätte.

Eine häufig zu hörende Fehlinterpretation des Signifikanztests ist es, aus der Zurückweisung der Nullhypothese H_0 auf einem bestimmten α^*-Niveau zu folgern, die Alternativhypothese H_1 sei mit einer Wahrscheinlichkeit $p = 1-\alpha^*$ wahr. Diese Interpretation mißachtet den Tatbestand, daß die empirische Signifikanz nur unter der Voraussetzung ermittelt wird, daß H_0 wahr ist. Nur unter dieser Voraussetzung kann man die

---

[4] Die Konzeption eines nachträglich festgestellten (»empirischen«) Fehlerrisikos ist wissenschaftstheoretisch umstritten. Insofern ist es auch umstritten, ob man beim zweiseitigen Test das a-priori Fehlerrisiko mit $\alpha \leqq p$ angeben kann, oder mit $\alpha = p$ angeben muß.

empirische Signifikanz bezüglich des realisierten Ereignisses überhaupt bestimmen. Wenn diese Wahrscheinlichkeit sehr gering ist, z. B. $\alpha^* \leqq 0{,}05$, hat man zwei Möglichkeiten der Interpretation: Man kann erstens annehmen, daß H_0 wahr ist und, wie der Zufall so mitspielt, ein sehr unwahrscheinliches Ereignis realisiert wurde. Oder man kann davon ausgehen, daß kein unwahrscheinliches Ereignis realisiert wurde, sondern H_0 falsch ist. Das Fehlerrisiko α^* bedeutet also nicht, daß H_0 mit der geringen Wahrscheinlichkeit $p = \alpha^*$ richtig wäre, folglich auch nicht, daß H_0 mit $p = 1-\alpha^*$ falsch wäre und somit auch nicht, daß die Negation H_1 mit $p = 1-\alpha^*$ richtig wäre.

Eine oft gegen den Signifikanztest vorgebrachte Kritik behauptet, er sei wertlos oder gar irreführend, da das Testergebnis von der manipulierbaren Größe des Stichprobenumfanges abhänge. Tatsächlich ist die empirische Signifikanz eines realisierten Stichprobenergebnisses durch den Stichprobenumfang mit bestimmt. In unserem Beispiel war eine Mittelwertdifferenz von | 1,12 | Jahren in der durchschnittlichen Mandatsdauer von adligen gegenüber nicht-adligen Reichstagsabgeordneten »nicht signifikant«, führte nicht zur Ablehnung der Nullhypothese. Wäre die gleiche Differenz in einer Stichprobe des Umfangs $n = 500$ beobachtet worden, hätte die Nullhypothese zurückgewiesen werden können (siehe Gleichung (8-12")). Doch dieser Tatbestand macht den Signifikanztest weder wertlos noch irreführend; er weist lediglich darauf hin, daß größere Stichprobenumfänge im allgemeinen zu kleineren Standardfehlern führen. Folglich braucht man bei kleineren Stichprobenumfängen größere Abweichungen vom hypothetisch angenommenen Wert, um die Nullhypothese mit einem gegebenen Fehlerrisiko zurückzuweisen. Aber die Zurückweisung der Nullhypothese impliziert keine Aussage über mögliche Werte des Populationsparameters, die nicht in der Nullhypothese formuliert worden sind. Eine »hohe« Signifikanz (kleines Fehlerrisiko) bedeutet z. B. nicht, daß der betreffende Parameter (z. B. ein Korrelationskoeffizient) in der Population einen hohen Wert hat. Aussagen hierüber müssen sich auf eine spezifische Intervallschätzung stützen. In sie gehen zwar teilweise die gleichen Faktoren ein wie bei der Signifikanzberechnung. »Stärke« und »(statistische) Signifikanz« einer Variablenbeziehung sind dennoch analytisch strikt auseinanderzuhalten. In die Irre führen kann also lediglich eine falsche Interpretation des Signifikanztests. Und daß das Fehlerrisiko bei sonst gleichen Beobachtungen mit dem Stichprobenumfang variiert, macht die Information über das jeweils gegebene Fehlerrisiko ja nicht wertlos.

8.5.3 »Stärke« eines Tests

Als Beta-Fehler haben wir im vorigen Abschnitt jene Fehlentscheidung bezeichnet, durch die eine Nullhypothese H_0 beibehalten wird, obwohl die

Alternativhypothese richtig ist. Wenn die Wahrscheinlichkeit hierfür mit $p = \beta$ angegeben ist, so gibt der Ausdruck $p = (1-\beta)$ an, mit welcher Wahrscheinlichkeit eine richtige Alternativhypothese durch einen Test auch als richtig aufgedeckt werden kann (siehe erneut Abb. 8.5). Diese durch die Größe $1-\beta$ abgestufte Fähigkeit eines Tests nennt man seine »Stärke« oder »Trennschärfe«.

Wir haben im vorigen Abschnitt gesehen, daß die Wahrscheinlichkeit β (und damit auch $1-\beta$) nur unter der Voraussetzung bestimmt werden könnte, daß der Populationsparameter bekannt wäre. Dem Statistiker ist es aber möglich, für ein bestimmtes Testverfahren die Wahrscheinlichkeit β oder $1-\beta$ in Abhängigkeit von einem variierenden Populationsparameter darzustellen. Die so ermittelte »Teststärkefunktion« kann als Entscheidungskriterium benutzt werden, wenn zur Überprüfung einer Hypothese mehrere statistische Tests zur Verfügung stehen.

Die Stärke eines Tests nimmt mit dem Umfang der Stichprobe zu, auf die er angewandt wird. Man kann also die Stärke unterschiedlicher Tests nur in bezug auf einen konstanten Stichprobenumfang miteinander vergleichen. Bei verschiedenen Tests nimmt die Stärke bei gleicher Erweiterung des Stichprobenumfangs in unterschiedlichem Maße zu, d. h., sie sind unterschiedlich »**effizient**«. Das Verhältnis zwischen Stärke und Effizienz wird als »**Stärke-Effizienz**« (»power-efficiency«) bezeichnet. Die entsprechenden statistischen Untersuchungen zeigen, daß Tests, die mit (informationsreichen) metrischen Daten arbeiten, grundsätzlich stärker sind als Tests, die mit Daten auf niedrigerem Meßniveau arbeiten. Allerdings muß man darauf achten, daß das vorausgesetzte Verteilungsmodell (z. B. Normalverteilung) auch tatsächlich erfüllt und das erforderliche Meßniveau gegeben ist.

8.5.4 Einseitige und zweiseitige Hypothesen

In unserem Beispiel zur Mandatsdauer der Reichstagsabgeordneten haben wir zwei Fassungen der Alternativhypothese vorgelegt:

(8-36) Fassung A: $H_1: \mu_1 \neq \mu_2$

Fassung B: $H_1: \mu_1 > \mu_2$

Man nennt die Fassung A eine »zweiseitige« Hypothese, die Fassung B eine »einseitige« Hypothese. Die Negation und damit die Nullhypothese zur Fassung A lautet

(8-37) $H_0: \mu_1 = \mu_2$ ((8-32) wiederholt)

woraus folgt: $E(\bar{x}_1 - \bar{x}_2) = 0$. Diese Hypothese ist widerlegbar durch extreme

Stichprobenergebnisse auf **beiden** Seiten der Verteilungsfunktion (siehe erneut Abb. 8.4).
Die Nullhypothese zur einseitigen Alternativhypothese (Fassung B) lautet

(8-38) $H_0: \mu_1 \leq \mu_2$ bzw. $E(\bar{x}_1 - \bar{x}_2) \leq 0$

Diese Nullhypothese kann also nur abgelehnt werden, wenn ein extremes Stichprobenergebnis auf **einer**, der rechten Seite der Verteilungsfunktion auftritt. Der kritische Wert (das Ausmaß der Extremität), der erreicht werden muß (wenn man H_0 ablehnen will) ist bei einem gegebenen Fehlerrisiko α betragsmäßig entsprechend geringer als bei der zweiseitigen Hypothese.

In unserem konkreten Beispiel zur Mandatsdauer der Abgeordneten hatten wir lediglich die zweiseitige Alternativhypothese bzw. die dazugehörige spezifische Nullhypothese (8-37) getestet. Das mit 100 Fällen erzielte Stichprobenergebnis von $(\bar{x}_1 - \bar{x}_2) = -1{,}12$ berechtigte nicht zur Zurückweisung der Nullhypothese. Wie können wir nun auch die unspezifische Nullhypothese (8-38) testen, nachdem wir bisher betont haben, daß sich (bei den üblichen Signifikanztests) nur spezifische Hypothesen testen lassen? Die Antwort besteht darin zu zeigen, daß ein Test der spezifischen Hypothese (8-37) gleichzeitig ein Test der unspezifischen Hypothese (8-38) ist. Das läßt sich mit Hilfe der Abbildung 8.6 demonstrieren.

In unserem konkreten Beispiel ist sofort deutlich: Wenn $\bar{x}_1 - \bar{x}_2 = -1{,}12$ nicht »ausreicht«, um die Hypothese H_0: $\mu_1 - \mu_2 = 0$ zurückzuweisen, dann wird dieses Ergebnis erst recht nicht ausreichen, die Hypothese $H_0^*(\mu_1 - \mu_2) < 0$ zurückzuweisen. Als Beispiel haben wir in Abb. 8.6 eine zweite Dichtefunktion für H_0^* $(\mu_1 - \mu_2) = a < 0$ eingezeichnet. Betrachten wir nun ein weiteres Stichprobenergebnis $\bar{x}_1 - \bar{x}_2 = +2{,}20$ bei unveränderter Standardabweichung $\sigma_{\bar{x}(1)-\bar{x}(2)} = 1{,}30$ (siehe Gleichung (8-11)). Dieses Ergebnis liegt um $2{,}20/1{,}30 = 1{,}69$ Standardabweichungen vom Erwartungswert Null entfernt. Es würde dazu berechtigen, die Nullhypothese H_0: $\mu_1 - \mu_2 = 0$, verstanden als partielle Negation der **einseitigen** Alternativhypothese in (8-36), zurückzuweisen: Einem z-Wert von $+1{,}69$ entspricht ein $\alpha < 0{,}05$. Erst recht kann dann eine Nullhypothese H_0^*: $\mu_1 - \mu_2 < 0$ zurückgewiesen werden, weil der beobachtete Wert vom Erwartungswert $(\mu_1 - \mu_2) = a < 0$ noch weiter entfernt liegt. Eine Ablehnung von H_0 impliziert also stets auch eine Zurückweisung von H_0^*. Das gilt aber nicht umgekehrt: Ein Stichprobenergebnis, das zur Zurückweisung von H_0^* führen würde, müßte nicht unbedingt auch die Zurückweisung von H_0 nach sich ziehen. Das schafft aber keine Probleme, denn in diesem Falle dürfte die Alternativhypothese erst recht nicht akzeptiert werden. Sie darf ja (bei konstanten Entscheidungsregeln) nur angenommen werden, wenn auch die H_0: $\mu_1 - \mu_2 = 0$ mit einem geringen Fehlerrisiko zurückgewiesen werden kann. Somit gilt allgemein, daß eine einseitige Alternativhypothese,

die zunächst zu einer unspezifischen Negation führt, auf indirektem Wege getestet werden kann, indem man die unspezifische Negation auf eine spezifische verkürzt und diese einem Test aussetzt.

8.6 Nicht-parametrische und verteilungsfreie Testverfahren (*)

Nehmen wir an, auf der Basis von Zensusbögen sei eine Stichprobe von 510 Haushalten nach dem Zufallsprinzip (siehe Kap. 9) ausgewählt worden. Die Haushalte können auf Grund der ergiebigen Informationen der Zensusbögen u. a. 6 Statuskategorien und 2 Mobilitätskategorien (hohe oder niedrige regionale Mobilität) zugeordnet werden. Bei der Statusvariable handelt es sich um eine Ordinalskala (mit gruppierten Daten). Die beiden Mobilitätskategorien definieren zwei Teilstichproben der Haushalte (oder auch Haushaltsvorstände). Sie sind unabhängig voneinander, sofern die Gesamtstichprobe nach dem Zufallsprinzip gezogen wurde. Die bivariate Häufigkeitstabelle mit den absoluten (f) und den kumulierten relativen Häufigkeiten (F) sieht so aus:

Status	niedrige Mobilität		hohe Mobilität		Differenz $F_1 - F_2$
	f_1	F_1	f_2	F_2	
1	58	.246	31	.113	.133
2	51	.462	46	.281	.181
3	47	.661	53	.474	.187
4	44	.847	73	.741	.106
5	22	.941	51	.927	.014
6	14	1.000	20	1.000	
	236		274		

Untersucht werden soll, ob das Mobilitätsniveau mit dem sozialen Status »zusammenhängt«. Eine solche Frage oder Hypothese haben wir bisher bei den statistischen Tests »übersetzt« in eine Annahme über den Wert eines bestimmten Parameters der Populationsverteilung, z. B. in eine Annahme über die Differenz der Mittelwerte in zwei Teilpopulationen. Als Nullhypothese haben wir z. B. formuliert: H_0: $\mu_1 - \mu_2 = 0$. Da in unserem fiktiven Beispiel keine metrischen Daten vorliegen, können wir diesmal keinen Test auf die »Zufälligkeit« einer Mittelwertdifferenz durchführen.

Einen Ausweg bietet ein »nicht-parametrisches« Testverfahren, das Kolmogorov und Smirnov vorgeschagen haben. In ihm wird die Nullhypothese, es bestehe kein Zusammenhang, in die Annahme übersetzt, daß die beiden Teilpopulationen (die Haushalte mit niedriger und die Haushalte mit hoher Mobilität) die gleiche Häufigkeitsverteilung des sozialen Status aufweisen. Eine Möglichkeit, zwei Verteilungen miteinander zu vergleichen, ohne irgendwelche Parameter zu spezifizieren, besteht darin, die kumulierten Häufigkeiten gegenüberzustellen und die maximal auftretende Differenz $|D_{max}|$ zu notieren. Je größer der Absolutbetrag der maximalen Differenz, umso größer ist der Unterschied zwischen den beiden Verteilungen. Dabei werden Unterschiede jedweder Art berücksichtigt, der Lokation ebenso wie z. B. der Streuung oder der Schiefe. Deshalb spricht man auch von einem »Omnibus-Test«. Den Statistikern war es möglich, die Stichprobenverteilung (sampling distribution) der Prüfgröße $|D_{max}|$ zu bestimmen, ohne irgendwelche Annahmen über die spezifische Verteilung der betreffenden Variable in der Grundgesamtheit (Population) zu machen. Wir wollen diese Stichprobenverteilung hier nicht in ihrer mathematischen Form angeben, sondern nur die daraus entwickelte Entscheidungsregel für die Testpraxis mitteilen. Der Test enthält unterschiedliche Fassungen, je nachdem, ob es sich a) um »große »oder »kleine« Stichproben oder b) um eine ein- oder eine zweiseitige Fragestellung handelt. Bei der einseitigen Fragestellung wird mit der Annahme (Forschungshypothese) operiert, daß die Gruppe A (z. B. die mit hoher Mobilität) in der Population insgesamt höhere Werte der Zufallsvariable (hier sozialer Status) aufweist, als die Gruppe B. Bei einem zweiseitigen Test wird keine Annahme über die Richtung des Zusammenhangs gemacht. Wenn die Teilstichproben Umfänge von $n_1 > 40$ und $n_2 > 40$ aufweisen (sie müssen nicht gleich groß sein), sieht die Entscheidungsregel bei einem zweiseitigen Test für die Ablehnung oder Beibehaltung der Nullhypothese wie folgt aus:
Wenn die maximale Differenz $|D_{max}|$ den kritischen Wert

$$D_\alpha = K_\alpha \frac{n_1+n_2}{n_1 \cdot n_2}$$

überschreitet, wird die Nullhypothese (die beiden Populationsverteilungen sind gleich) mit dem Fehlerrisiko α zurückgewiesen; ist $|D_{max}| \leqq D_\alpha$ wird sie beibehalten. Der Wert K_α hängt nur von dem gewählten Signifikanzniveau ab. Siegel (1956, S. 279) gibt die K_α-Werte für die gebräuchlichsten α-Niveaus.

α	.10	.05	.025	.01	.005	.001
K_α	1.22	1.36	1.48	1.63	1.73	1.95

Wenn wir für unser Testbeispiel ein Fehlerrisiko von $\alpha \leqq 0.05$ wählen, müssen wir also mit einem $K_\alpha = 1.36$ operieren. Der kritische Wert D_α ergibt sich somit aus:

$$D_\alpha = 1.36 \frac{236+274}{236 \cdot 274} = 0.121$$

Der Tabelle mit den kumulierten Häufigkeiten entnehmen wir eine beobachtete maximale Differenz von $|D_{max}| = .187$. Sie ist größer als D_α. Somit können wir die Nullhypothese mit einem Fehlerrisiko von $\alpha \leqq 0.05$ zurückweisen.

Wie der Test bei einseitiger Fragestellung oder bei kleineren Stichprobenumfängen durchzuführen ist, erläutert Siegel (1956, S. 127ff.). Der Test kann auch angewandt werden, wenn eine empirische mit einer (vorgegebenen) theoretischen Verteilung verglichen werden soll.

Stellen wir noch einmal die zwei Besonderheiten dieses Testverfahrens heraus: (a) Es kann angewandt werden, ohne daß man Annahmen über die Verteilung der betreffenden Variablen in der Grundgesamtheit machen muß. Es entfiel z. B. die Annahme, die Zufallsvariable sei in der Population normalverteilt. Insofern kann man von einem »verteilungsfreien »Testverfahren sprechen[5]. (b) Getestet wurde kein hypothetisch angenommener Wert eines »Parameters«, wie z. B. eines bestimmten Mittelwertes oder einer Mittelwertdifferenz, sondern die »globale« Unterschiedlichkeit zweier Verteilungen. Insofern läßt sich das Verfahren als »parameterfrei« oder »nicht-parametrisch« kennzeichnen.

Die in der Literatur häufig zu findende Gleichsetzung zwischen »verteilungsfreien« und »nichtparametrischen« Verfahren ist jedoch nicht unproblematisch. Der Signifikanztest auf die »Zufälligkeit« einer Mittel-

[5] Das Attribut »verteilungsfrei« sollte nicht zu dem Mißverständnis führen, der Test käme ohne eine präzise definierte Stichprobenverteilung der jeweiligen Prüfgröße aus.

wertdifferenz ($\bar{x}_1 - \bar{x}_2$) ist sicherlich ein Parametertest. Bei seiner Durchführung formulieren wir (in der Regel) die Nullhypothese, der Populationsparameter ($\mu_1 - \mu_2$) sei gleich Null. Wenn ausreichend große Stichproben vorliegen, werden aber zur Ableitung der Stichprobenverteilung für die Statistik ($\bar{x}_1 - \bar{x}_2$) keine Annahmen über die Verteilung von X in der Population benötigt. Insofern ist dieser Test »verteilungsfrei«.

In den meisten Lehrbüchern wird auch der im nächsten Kapitel zu besprechende χ^2-Test der Rubrik der verteilungsfreien Testverfahren zugeordnet. Diese Etikettierung ist aber fragwürdig, wenn der χ^2-Test als Unabhängigkeitstest durchgeführt wird. Deshalb haben wir ihn nicht in diesem Abschnitt erläutert.

Nicht-parametrische bzw. verteilungsfreie Tests sind nicht nur für nicht-metrische Variablen konstruiert worden. Sie stellen auch für metrische Variablen eine Alternative zu den parametrischen bzw. verteilungsgebundenen Testverfahren dar, falls die u.U. sehr restriktiven Annahmen über die Populationsverteilung nicht realistisch sind. Allerdings sind die verteilungsfreien (nicht-parametrischen) Verfahren weniger trennscharf, implizieren also bei gleicher Stichprobengröße und gleichem α-Niveau ein höheres Risiko, bei Beibehaltung der Nullhypothese einen Fehler vom Typ II (siehe Abschn. 8.5.2 und 8.5.3 zu begehen).

Einen leicht lesbaren Überblick zu den verschiedenen »nicht-parametrischen« Verfahren gibt das Standardwerk von Siegel (1956). Eine Vielzahl dieser Tests steht in SPSS^x zur Verfügung.

8.7 Weitere Anwendungsbeispiele zu einzelnen Testverfahren

8.7.1 Anteilsdifferenzen (Differenzen von Proportionen)

Wir haben schon an anderer Stelle darauf hingewiesen, daß ein Anteilswert p als arithmetisches Mittel der Häufigkeitsverteilung einer binär kodierten Variablen X aufzufassen ist, deren eine Merkmalsausprägung den Wert »1« und deren andere den Wert »0« erhält:

$$p = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (8\text{-}39)$$

X sei die Variable »Religionsbekenntnis der Reichstagsabgeordneten von 1912«. Mit X = 1 wird ein protestantisches Bekenntnis, mit X = 0 werden alle anderen Bekenntnisformen kodiert. Von 462 erfaßten Reichstagsab-

geordneten des Jahres 1912 sind 212 protestantisch und 250 nicht protestantisch (siehe Abb. 8.7). Wir erhalten also

$$(8\text{-}40)\quad p(prot) = (212 \cdot 1 + 250 \cdot 0)/462 = 212/462 = 0{,}459$$

Wie wir Abb. 8.7 entnehmen können, ergeben sich für bürgerliche und adlige Abgeordnete hinsichtlich des Religionsbekenntnisses unterschiedliche Verteilungen: Von den bürgerlichen Abgeordneten sind 43,4 % protestantisch ($p_1 = 0{,}434$), von den adligen Abgeordneten 60,6 % ($p_2 = 0{,}606$). Folglich beobachten wir eine Anteilsdifferenz von $p_2 - p_1 = 0{,}173$.

Sofern wir diese Differenz und damit einen »Zusammenhang« zwischen »Stand« und »Religionsbekenntnis« nur für diese Gruppe von Reichstagsabgeordneten behaupten wollen, ergibt sich keine Notwendigkeit für einen statistischen Signifikanztest. Der Übung wegen wollen wir jedoch die 462 Reichstagsabgeordneten als Zufallsstichprobe aus einer Population auffassen.

Falls wir a priori einen höheren Protestantenanteil bei den Adligen erwartet haben, ist unsere Forschungshypothese

$$H_1: \pi_2 - \pi_1 > 0,$$

die Nullhypothese folglich:

$$H_0: \pi_2 - \pi_1 \leqq 0.$$

Wir führen also einen einseitigen Test durch. Da wir es mit relativ großen Teilstichproben ($n_1 = 396$ bürgerlichen und $n_2 = 66$ adligen Abgeordneten) zu tun haben und $n_1 \cdot p_1 > 10$, $n_2 \cdot p_2 > 10$ ist (siehe oben, S. 45), können wir in diesem Falle auch für die Anteilsdifferenz (analog zur Mittelwertdifferenz) das Normalverteilungsmodell unterstellen. Die Anteilsdifferenz $p_2 - p_1$ streut bei wiederholter Stichprobenziehung um die Populationsdifferenz $\pi_2 - \pi_1$ mit einer Standardabweichung

$$(8\text{-}41)\qquad \sigma_{p(1)-p(2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

(siehe Gleichung (8-8))

Da normalerweise π_1 und π_2 und damit auch $\sigma_{P_2-P_1}$ in einer Testsituation nicht bekannt sind, müssen diese Kennzahlen anhand der Stichprobendaten geschätzt werden. Es läßt sich zeigen, daß der Stichprobenanteil $p = \hat{\pi}$ ein erwartungstreuer Schätzer des Populationsanteils π ist. Seine Varianz ist mit

$$(8\text{-}42) \qquad \sigma_p^2 = \frac{\pi(1-\pi)}{n}$$

gegeben. Sie kann - ebenfalls erwartungstreu - mit der entsprechenden Stichprobenvarianz p(1-p) geschätzt werden, so daß wir für den Standardfehler in (8-41) den Schätzer

$$(8\text{-}43) \qquad \sigma_{p(1)-p(2)} = s_{p(1)-p(2)} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

erhalten. Freiheitsgrade gehen hierbei nicht verloren, da man die Standardabweichung direkt aus den Anteilswerten ermittelt. Damit sind alle Elemente versammelt, die wir zur Bestimmung der Prüfgröße z der Standardnormalverteilung benötigen:

$$(8\text{-}44) \quad z = \frac{p_2 - p_1 - E(p_2 - p_1)}{s_{p(1)-p(2)}} = \frac{0{,}172 - 0}{\sqrt{\frac{0{,}434 \cdot 0{,}566}{396} + \frac{0{,}606 \cdot 0{,}394}{66}}}$$

$$z = \frac{0{,}172}{0{,}065} \approx 2{,}63$$

Bei einem einseitigen Test und einem Alpha-Niveau von $\alpha = 0{,}05$ liegt der kritische z-Wert bei $z = |1{,}645|$. Jeder größere z-Wert bedeutet also ein empirisches Fehlerrisiko von $\alpha^* < 0{,}05$ für die Zurückweisung der Nullhypothese. In unserem Beispiel mit $z = |2{,}63|$ ist $\alpha^* \approx 0{,}0043$. Die Nullhypothese kann also mit einem sehr geringen Fehlerrisiko zurückgewiesen werden. (Es sei an die Ausführungen in Abschnitt 8.5.4 erinnert, wonach die Zurückweisung von H_0: $\pi_1 - \pi_2 = 0$ die Zurückweisung von H_0^*: $\pi_1 - \pi_2 < 0$ impliziert).

Im allgemeinen kann man sich diesen etwas umständlichen Test ersparen und statt dessen den Chi-Quadrat-Test bzw. den Fisher Exact Test (siehe den folgenden Abschnitt) anwenden, die bei der Tabellenanalyse mit SPSS über das STATISTICS-Kommando abgerufen werden können. Im Falle kleiner Stichproben, wenn also das Normalverteilungsmodell nicht unterstellt werden kann, ist der Fisher Exact Test vorzuziehen (Blalock 1960, 177).

8.7.2 Der Chi-Quadrat-Unabhängigkeitstest

Wir haben in Abschnitt Teil I, 4.2.2 einige Zusammenhangsmaße kennengelernt, die auf der Größe χ^2 beruhen, wie z. B. den Kontingenzkoeffizienten C oder Cramers V. Diese Kennzahlen drücken die »Stärke« eines beobachteten Zusammenhangs aus, aber nicht unmittelbar dessen »Signifikanz«. Die Signifikanz läßt sich jedoch direkt über die χ^2-Größe testen, wie sie in Teil I, Gleichung (4-7) definiert ist. Unter der Nullhypothese, daß die beiden Variablen stochastisch unabhängig sind, folgt die Verteilung von χ^2 näherungsweise der Chi-Quadrat-Verteilung, wie wir sie in Abschn. 7.4.1 vorgestellt haben. Die Freiheitsgrade (df) ergeben sich aus der Anzahl der Spalten, c, und der Zeilen, r, der entsprechenden Tabelle:

$$df = (r-1)\,(c-1)$$

Für die bivariate Verteilung[6] in Abb. 8.7 (4-Felder-Tafel), mit der wir im vorigen Abschnitt den Test auf die Signifikanz einer Anteilsdifferenz erläutert haben, erhalten wir ein $\chi^2 = 6{,}72$ mit $(2-1)\,(2-1) = 1$ Freiheitsgrad. Der SPSS-Ergebnisausdruck weist hierfür eine Signifikanz von $\alpha^* = 0{,}0095$ aus. Dieser Wert entspricht bis auf Rundungsfehler dem doppelten Fehlerrisiko $\alpha^* = 0{,}0043$, das wir im vorigen Abschnitt beim **einseitigen** Test auf die Signifikanz der Differenz im Protestantenanteil bürgerlicher und adliger Abgeordneter ermittelt hatten. Wenn bei dem einen Test das Normalverteilungsmodell Z und beim anderen das Chi-Quadrat-Modell (für 1 Freiheitsgrad) unterstellt werden kann, gilt $Z_{(1-\alpha/2)} = \sqrt{\chi^2_{1-\alpha}}$, in unserem Beispiel (siehe 8-44): $2{,}63 \approx \sqrt{6{,}72}$. Das Fehlerrisiko für die Zurückweisung der Nullhypothese wird um so niedriger, je größer der Wert χ^2 (oder z).

Die Appoximation an die Chi-Quadrat-Verteilung gilt im allgemeinen als hinreichend genau, wenn in allen Zellen der Tabelle die **erwartete** Häufigkeit größer/gleich 5 ist. In der sozialwissenschaftlichen Forschungspraxis folgt man aber häufig einer etwas »weicheren« Regel: Der Test wird lediglich dann als nicht anwendbar betrachtet, wenn mehr als ein Fünftel aller Zellen eine erwartete Häufigkeit von kleiner als 5 haben oder wenn eine der erwarteten Häufigkeiten kleiner als 1 ist. In der (3x3)-Tabelle in Abb. 4.9 (Teil I) sind zwar zwei Zellen überhaupt nicht besetzt, aber die erwarteten Häufigkeiten (siehe Abb. 4.10) sind auch für diese Zellen größer als 5. Wir können also in diesem Beispiel den Chi-Quadrat-Test auch

[6] Bei tri- und mehrvariaten Verteilungen (den entsprechenden mehrdimensionalen Tabellen) läßt sich auch die stochastische Unabhängigkeit von drei oder mehr Variablen testen. Die Erwartungswerte für die Besetzung der einzelnen Zellen werden nach dem gleichen Prinzip berechnet. Wenn $K_1, K_2, \ldots, K_m$ die Zahl der Kategorien der ersten, zweiten, ..., m-ten Variablen bezeichnen, ergibt sich die Zahl der Freiheitsgrade aus $df = (K_1 \cdot K_2 \cdot \ldots \cdot K_m - 1) - [(K_1-1)+(K_2-1)+\ldots+(K_m-1)]$.

nach der strengeren Regel anwenden: Der Wert $\chi^2 = 19{,}80$ (siehe Gleichung (4-7)) führt bei (3-1) (3-1) = 4 Freiheitsgraden zu einem Fehlerrisiko von $\alpha < 0{,}001$.

Die summarische Prüfgröße χ^2 sagt über die Natur des Zusammenhangs der beiden Variablen wenig aus, sie informiert nur über die statistische Abhängigkeit/Unabhängigkeit der beiden Variablen. Das Ergebnis ist, wie wir in Teil I sahen, abhängig von der Größe der Stichprobe. Aussagekräftiger sind die Assoziationsmaße (wie Cramer's V), die unabhängig von der Fallzahl berechnet werden. Inhaltlich noch aufschlußreicher sind oft die Differenzen zwischen den beobachteten und den erwarteten Häufigkeiten in den einzelnen Zellen, die sog. »Residuen«. Es empfiehlt sich, diese Residuen nicht in ihren Rohwerten sondern in »standardisierten« Größen zu ermitteln. Üblicherweise werden sie standardisiert, indem man die Differenzenbeträge zur Wurzel der jeweils erwarteten Häufigkeit ins Verhältnis setzt, d. h., man zieht die Wurzel aus den Komponenten, die in ihrer Summe die Prüfgröße χ^2 bilden (vergl. (4-7)):

$$(8\text{-}45) \qquad \text{stand. Residuen} = \frac{f_b - f_e}{\sqrt{f_e}} = \sqrt{\frac{(f_b - f_e)^2}{f_e}}$$

Im Falle von (2x2)-Tabellen versucht man die Approximation an die Chi-Quadrat-Verteilung durch die sog. Kontinuitätskorrektur von Yates zu verbessern. Eine entsprechend korrigierte χ^2-Größe wird in SPSS automatisch mitgeliefert. Bei sehr kleinen Fallzahlen (Daumenregel: $n \leqq 20$) wird statt des Chi-Quadrat-Tests der Fisher Exact Test durchgeführt, der auf einem hypergeometrischen Verteilungsmodell beruht[7].

Man kann den Chi-Quadrat-Test in einem anderen Kontext auch als sog. **Anpassungstest** verwenden, wenn geprüft werden soll, ob ein theoretisches Verteilungsmodell für diskrete Daten oder klassierte Intervalldaten adäquat an eine empirische Häufigkeitsverteilung angepaßt ist. Auf diese Weise läßt sich z. B. das Normalverteilungsmodell mit einer beobachteten Häufigkeitsverteilung vergleichen und bewerten. Dazu rechnet man für bestimmte Quantilsabstände die Häufigkeiten aus, die nach dem Normalverteilungsmodell zu erwarten sind und vergleicht sie mit den beobachteten Häufigkeiten.

Zu erwähnen ist noch, daß sich unter der Voraussetzung $\chi^2 \neq 0$ auch Standardfehler für die Stichproben-Assoziationsmaße angeben lassen, die eine Funktion von χ^2 sind (siehe Teil I, Kap. 4.2.2). Man findet die ent-

---

[7] Dieser Test ist außer in Siegel (1956) beispielsweise auch in Blalock (1960, S. 220 ff.) näher beschrieben.

sprechenden Angaben z. B. in Hartung et al. (1986, S. 452). Mit diesen Informationen lassen sich die entsprechenden Konfidenzintervalle schätzen.

8.7.3 Test auf Signifikanz des Pearsonschen Korrelationskoeffizienten r(*)

Will man Pearsons Korrelationskoeffizienten r (siehe Teil I, Abschn. 4.2.4) auf seine Signifikanz testen oder Konfidenzintervalle für den entsprechenden Populationsparameter Rho (ρ) schätzen, unterstellt man üblicherweise ein bivariates Normalverteilungsmodell (siehe Abb. 8.8) für die beiden Variablen X und Y. Die Merkmale X und Y müssen also nicht nur jedes für sich normalverteilt sein, sondern es müssen auch die zu einem Wert $X=x_i$ ($Y=y_i$) gehörenden Verteilungen der Y-Werte (X-Werte) normalverteilt sein. Außerdem wird vorausgesetzt, daß die Varianzen der bedingten Verteilungen gleich sind. (Zu den Konsequenzen, die sich bei Verletzung dieser Voraussetzungen ergeben, findet man Literaturhinweise in Bortz 1979, S. 259). Aber die Stichprobenkorrelation r ist selbst bei Erfüllung dieser Voraussetzungen kein erwartungstreuer, sondern lediglich ein konsistenter Schätzer für die Populationskorrelation ß.

Die Stichprobenverteilung von r hängt davon ab, welche Korrelation ρ in der Population vorliegt. Bei $\rho = 0$ ist r bei hinreichend großem Stichprobenumfang annähernd normalverteilt (Bortz 1979, S.259). Bei kleinen Stichproben mit $n \geqq 4$ ist der transformierte r-Koeffizient

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}} \qquad (8\text{-}46)$$

annähernd t-verteilt mit n-2 Freiheitsgraden. Für den Signifikanztest können wir also je nach Stichprobengröße die bereits vertrauten z- und t-Tabellen heranziehen.

Häufiger allerdings wird die Stichprobenkorrelation indirekt im Rahmen des Regressionsmodells auf ihre Signifikanz getestet. Wie wir in Kap. 10 sehen werden, muß $\rho = 0$ sein, wenn der Regressionskoeffizient $ß = 0$ ist (und umgekehrt). Für den Test auf $ß = 0$ ist die Normalverteilungsvoraussetzung (bei großem Stichprobenumfang) weniger gravierend. Damit ist jedoch noch nicht das Problem gelöst, wie Konfidenzintervalle für den Korrelationskoeffizienten zu konstruieren sind, wenn $\rho \neq 0$ ist.

Der Statistiker R. A. Fisher konnte zeigen, daß sich die Korrelationskoeffizienten r so transformieren lassen, daß sie auch dann annähernd normalverteilt sind, wenn $\rho \neq 0$ ist. In der Literatur spricht man von »Fishers Z-Transformation«:

$$(8\text{-}47) \qquad Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

Diese Z-Werte sind mit einem Standardfehler von $\sigma_Z = \sqrt{1(n-3)}$ approximativ normalverteilt. (Sie dürfen nicht mit den z-Werten der Normalverteilungstabelle verwechselt werden.) Es empfielt sich , auch dann auf die Z-Transformation zurückzugreifen, wenn man annimmt, daß $\rho = 0$ ist).

In Teil I, Abschnitt 4.2.4 hatten wir für den Zusammenhang zwischen den Stimmenanteilen der SPD und den Anteilen der in Industrie und Gewerbe Beschäftigten der jeweiligen Wahlkreise eine Korrelation von r= 0,63 errechnet. Nehmen wir einmal an, es handele sich dabei um ein Stichprobenergebnis und wir wollten ein Konfidenzintervall für den Populationsparameter ρ schätzen. Viele Statistik-Lehrbücher enthalten Tabellen, in denen die r-Werte in Z-Werte umgerechnet worden sind (siehe Anhang A, Tab. 5). Für r = 0,63 ergibt sich ein Z = 0,741. Folglich erhalten wir das Konfidenzintervall mit

$$(8\text{-}48) \qquad Z \pm z_{0,05} \cdot \sqrt{1/(n-3)} = 0,741 \pm 1,96 \cdot \sqrt{1/(n-3)}$$

$$= 0,741 \pm 0,102$$

$$z_1 = 0,639 \quad , \quad z_2 = 0,843$$

Wiederum laut Umrechnungstabelle ergeben sich daraus die Intervallgrenzen

$$(8\text{-}49) \qquad (0,56 < \rho < 0,69)$$

Das Konfidenzintervall um r ist nicht mehr symmetrisch. Es weicht von der Symmetrie um so stärker ab, je kleiner n.

Im Gegensatz zu den r-Werten stellen die Z-Werte eine Ratioskala dar. Während ein $r_2 = 0,6$ gegenüber einem $r_1 = 0,3$ **keinen** »doppelt so starken« (linearen) Zusammenhang indiziert, läßt ein $Z_2 = 0,5$ im Vergleich zu einem $Z_1 = 0,25$ auf einen doppelt starken Zusammenhang schließen.

Wie man Korrelationskoeffizienten, die man in zwei oder mehreren voneinander unabhängigen Stichproben ermittelt hat, auf ihre Unterschiedlichkeit testen kann, erläutert z. B. Bortz (1979, S.263f.).

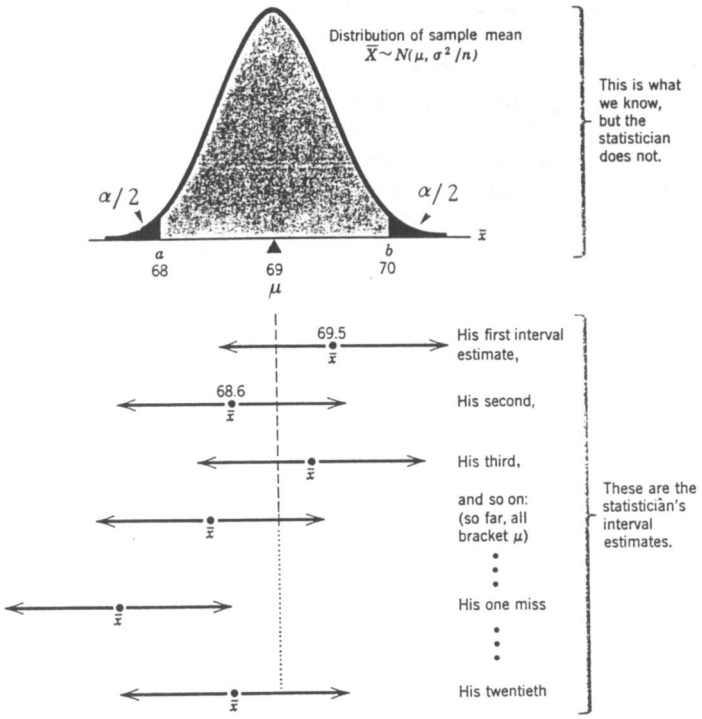
8.7.4 Signifikanztests für PRE-Maßzahlen

Die in Teil I, Abschnitt 4.2.3 vorgestellten PRE-Maße Gamma, Tau und Somers d sind alle unter dem multinomialen Verteilungsmodell (siehe Abschn. 7.2) asymptotisch (mit zunehmender Stichprobengröße) normalverteilt mit Varianzen, deren Definitionsgleichungen ziemlich umfänglich sind. Die Formeln hierfür sind z. B. in Liebetrau (1983) zu finden. Entscheidend für den Signifikanztest ist wiederum das Verhältnis des Stichprobenkoeffizienten zu seinem (geschätzten) Standardfehler.

Zur Konstruktion von Konfidenzintervallen (wenn also nicht vorausgesetzt werden kann, daß die Populationsmaßzahl gleich Null ist) müssen andere Standardfehler eingesetzt werden als beim Signifikanztest.

SPSS^x bietet den Signifikanztest lediglich für die Tau-Koeffizienten an, während das Programmpaket BMDP entsprechende Testergebnisse auch für die anderen PRE-Maßzahlen (und weitere Assoziationsmaße) liefert. Der BMDP-Ergebnisausdruck enthält auch die Standardfehler, die bei der Konstruktion von Konfidenzintervallen benötigt werden.

Abb. 8.1: Veranschaulichung der Logik des Schätzens von Intervallen (wiederholtes Schätzen bei 20 Stichprobenziehungen)



Quelle: Wonnacott/Wonnacott 1972, S. 145

Abb. 8.2: Dichtefunktionen verschiedener Schätzer für den Parameter θ

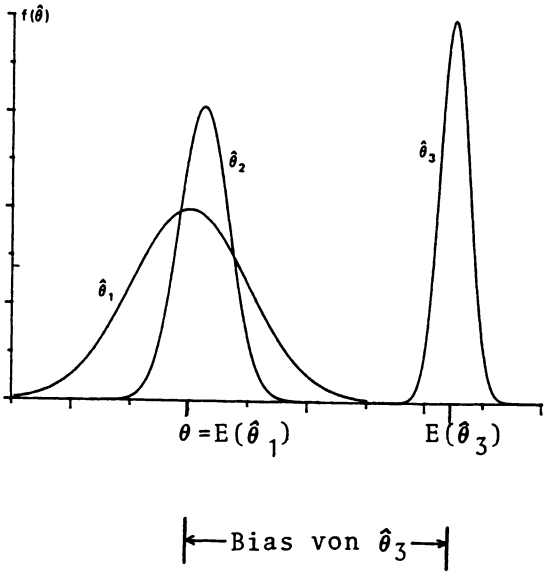
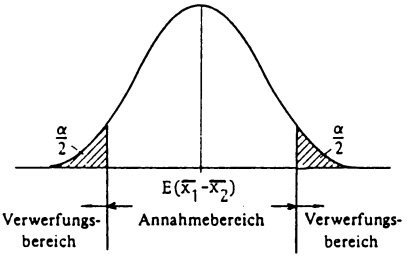


Abb. 8.3: Fehlertypen beim Signifikanztest

		In der Population gilt die:	
		H_0	H_1
Entscheidung auf Grund der Stichprobe zugunsten der:	H_0	richtige Entscheidung	β-Fehler
	H_1	α-Fehler	richtige Entscheidung

Quelle: Bortz 1979, S. 142

Abb. 8.4: Annahme- und Verwerfungsbereich beim Signifikanztest



Quelle: Bortz 1979, S. 150 (modifiziert)

Abb. 8.5: Abhängigkeit des ß-Fehlers von dem unbekannten Populationsparameter Theta

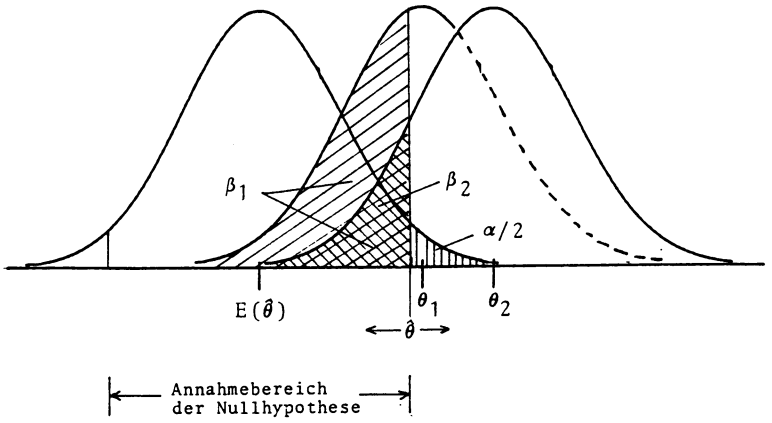


Abb. 8.6: Zur Veranschaulichung von Fehlerwahrscheinlichkeiten bei unspezifischer Nullhypothese

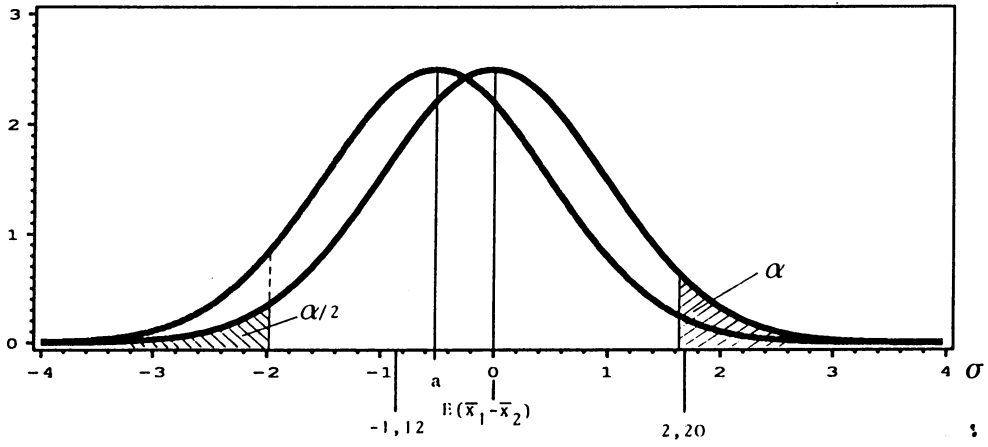
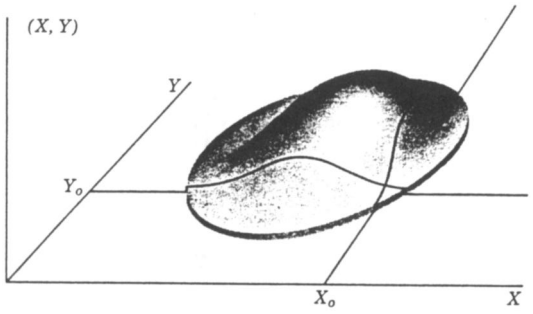


Abb. 8.7: Bivariate Verteilung von Konfession und "Stand" der Reichstagsabgeordneten von 1912 (Spaltenprozentuierung)

		Stand		
	COUNT I			
	COL PCT I	BUERGERL IICH	ADELIG	ROW TOTAL
		0I	1.I	
Konfession				
	1. I	172	40	212
PROTESTANTISCH	I	43.4	60.6	45.9
	2. I	224	26	250
ANDERE	I	56.6	39.4	54.1
	COLUMN	396	66	462
	TOTAL	85.7	14.3	100.0

Abb. 8.8: Bivariate Normalverteilung



Quelle: Wonnacott/Wonnacott 1972, S. 332

KAPITEL 9

Auswahlverfahren[1] (*)

9.1 Einleitende Bemerkungen

Die meisten Historiker sind an das Prinzip gewöhnt, **alle** verfügbaren Quellen auszuwerten, die zu einem bestimmten Untersuchungsgegenstand gehören und erreichbar sind. Es fällt jedoch schwer, und meistens ist es auch unsinnig, diesem Prinzip zu folgen, wenn Quellen über gleichartige Objekte vieltausend- oder gar millionenfach vorliegen. Wer z. B. über soziale Herkunft, Familienstatus, Gesundheitszustand und Karriereverlauf britischer Soldaten des ersten Weltkrieges forschen will, findet in dem entsprechenden Archiv des Verteidigungsministeriums Millionen von Personalakten vor. Es wäre viel zu aufwendig, **jede** einzelne Akte maschinenlesbar zu machen und auszuwerten. Sinnvoller ist es, eine »repräsentative« Auswahl von einigen tausend Fällen aus diesem Aktenbestand zu ziehen, wie das z. B. von Doron Lamm (1988) beschrieben worden ist (zum Begriff der Repräsentativität siehe unten). Wenn ein solches Auswahlverfahren bestimmten Regeln folgt, die wir gleich besprechen werden, lassen sich die Untersuchungsergebnisse, die sich auf die ausgewählten Fälle stützen, mit angebbaren Fehlerrisiken auf die Gesamtheit der Fälle »übertragen«.

Die Menge aller »Objekte«, über die man Aussagen machen will, nennt man eine **Grundgesamtheit** oder **Population**. Wir wollen für die Zwecke dieses Kapitels unterstellen, daß es sich dabei um eine empirische (nicht eine hypothetische) Population mit einer endlichen Zahl von Objekten handelt. Falls die Daten aller Elemente einer Grundgesamtheit erhoben werden, spricht man von einer **Vollerhebung**, andernfalls von einer **Teilerhebung**. Folgt man bei der Teilerhebung bestimmten Regeln (siehe unten), bezeichnet man sie als **Stichprobe** (»Sample«). Maßzahlen wie z. B. Mittelwerte oder Regressionskoeffizienten bezeichnet man als **Statistiken**, wenn sie für Stichproben berechnet werden, und als **Parameter**, wenn sie sich auf die Grundgesamtheit beziehen.

In der historischen Sozialforschung kommt es häufiger als in der empirischen Sozialforschung vor, daß die tatsächliche Auswahlgesamtheit (»**frame population**«), aus der die Stichprobe gezogen wird, erheblich von der angezielten Grundgesamtheit (»**target population**«) abweicht, über de-

[1] Dieses Kapitel orientiert sich stark an der Darstellung in Schnell/Hill/Esser 1988, Kap. 6.

ren Objekte man Aussagen machen möchte. So mag man z. B. Aussagen über **alle** Rekruten (die target population) anstreben, die zwischen 1912 und 1918 von der britischen Armee eingezogen wurden. Doch die Aktenbestände (frame population), aus denen man auswählen möchte, weisen Lücken auf, viele sind zerstört worden oder auf andere Weise verloren gegangen.

Eine Stichprobe kann man nur aus einer Gesamtheit ziehen, deren Elemente in irgendeiner Art von »Liste« erfaßt sind. Diese Liste kann nicht nur unvollständig sein (»undercoverage«); sie kann auch Fälle enthalten, die nicht zur Grundgesamtheit gehören (»overcoverage«) - z. B. Rekruten, die außerhalb des festgelegten Zeitraums eingezogen worden sind. Beides kann z. B. bei fehlerhaftem Kombinieren unterschiedlicher Datenquellen vorkommen. Weiterhin können Fehler dadurch auftreten, daß einige Fälle unerkannt mehrfach in der Liste enthalten sind. Andere gelangen vielleicht deshalb nicht in die Stichprobe, weil sie zwar nach Plan ausgewählt, aber aus irgendwelchen Gründen nicht erreichbar sind. Wieder andere fallen heraus, weil sie im Verlauf der Datenverarbeitung verloren gehen. Der tatsächlichen Stichprobe ordnet man konzeptuell die sog. **Inferenzpopulation** zu. Damit bezeichnet man diejenige Grundgesamtheit, für die die vorliegende Stichprobe tatsächlich eine (Zufalls-)Auswahl darstellt.

In der historischen Sozialforschung ist die Differenz zwischen angestrebter Grundgesamtheit einerseits und Auswahlgesamtheit oder Inferenzpopulation häufig nicht nur groß, sondern oft auch nicht genau angebbar: Man weiß nicht mit Sicherheit, wieviele Fälle verloren gegangen und mit welchen Anteilen sie welchen Kategorien zuzuordnen sind. (Zur allgemeinen Einführung in die Auswahlproblematik im Rahmen der historischen Sozialforschung siehe Floud 1980, Kap. 8 und Rohlinger 1982.). Um diese Situation zu kennzeichnen, spricht man auch von »sich selbst auswählenden« Stichproben (»**self selected samples**«), von Auswahlvorgängen, die der Forscher nicht gestalten und kontrollieren, u. U. aber modellhaft rekonstruieren kann. (Zur Beschreibung eines solchen Selektionsvorgangs siehe den Bericht von Shapiro et al. 1987.) Der inferenzstatistische Umgang mit self selected samples und rekonstruierten Selektionsvorgängen wirft natürlich erhebliche Probleme auf, die gesondert erörtert werden müßten (siehe hierzu Berk 1983). In diesem Kapitel beschränken wir uns auf die Darstellung verschiedener Typen von Auswahlverfahren, bei denen der Forscher selbst nach einem bestimmten Plan eine Stichprobe zieht. Nur wenn er dabei Regeln der Zufallsauswahl folgt, erhält er ein Sample, das eine (wahrscheinlichkeits-)theoretisch abgesicherte Anwendung inferenzstatistischer Methoden (siehe Kap. 6 bis 8) erlaubt. Nur für Zufallsstichproben (auch »Wahrscheinlichkeitsauswahlen« genannt) sind die Grenzen der Fehler, die beim Schluß von der Stichprobe

auf die Grundgesamtheit entstehen, exakt berechenbar. Das bedeutet nicht, daß aus anderen Stichproben keine korrekten Schlußfolgerungen gezogen werden können. Es bedeutet aber, daß deren Gültigkeit nicht gesichert ist und nicht anhand inferenzstatistischer Prinzipien getestet werden kann.

9.2 Die einfache Zufallsauswahl

Aus einer exakt definierten Grundgesamtheit mit N Elementen können $\binom{N}{n}$ verschiedene Stichproben mit n Elementen gezogen werden. Bei der einfachen Zufallsauswahl sind die Auswahlregeln so gestaltet, daß jede mögliche Stichprobe mit n, $n < N$ Elementen dieselbe Chance besitzt, gezogen zu werden. Dies impliziert, daß jedes einzelne Element der Grundgesamtheit die gleiche Chance hat, in die Stichprobe zu gelangen. Wird aus einer einfachen Zufallsstichprobe erneut eine einfache Zufallsauswahl vorgenommen, erhält man eine Stichprobe, die ebenfalls eine einfache Zufallsauswahl aus der Grundgesamtheit darstellt.

Einfache Zufallsstichproben werden in einem einzigen Auswahlvorgang aus einer Grundgesamtheit gezogen (im Gegensatz zu mehrstufigen Auswahlverfahren - siehe unten). Dazu gibt es verschiedene technische Möglichkeiten (siehe Böltken 1976). Wenn jedes Element der Grundgesamtheit eine Identifikationsnummer hat, lassen sich die Stichprobenelemente unter Verwendung sog. **Zufallszahlen** auswählen. Wie das im einzelnen vor sich geht, ist z. B. in Floud 1980, S. 180 f. und Blalock 1960, S. 395 beschrieben. Zufallszahlentabellen sind in vielen Statistik-Lehrbüchern (z. B. Blalock 1960) enthalten oder können mit dem Computer erzeugt werden. Zufallszahlen sind dadurch charakterisiert, daß ihre Abfolge keinerlei Systematik aufweist und die Häufigkeiten, mit denen die einzelnen Zahlen auftreten, gleich sind. Computeralgorithmen kommen diesem Ideal lediglich nahe; deshalb spricht man hier von »Pseudo-Zufallsgeneratoren«. Gebräuchlich sind auch sog. »systematische« Auswahlverfahren, bei denen z. B. »jede n-te Karte« gezogen wird, wobei jede Karte ein Element der Grundgesamtheit repräsentiert. Der Zufallscharakter dieser Auswahlverfahren ist umstritten (siehe Böltken 1976, S. 165 - 170). Er ist dann nicht gegeben, wenn die Anordnung der Fälle eine Periodizität (z. B. im Wechsel den männlichen und weiblichen Ehepartner) aufweist, die mit Untersuchungsvariablen korreliert.

Die Schätzfunktionen, Stichprobenverteilungen und Testverfahren, die wir in den vorangegangenen Kapiteln besprochen haben, sind alle unter der Voraussetzung abgeleitet worden, daß man es mit einfachen Zufallsstichproben zu tun hat. Die entsprechenden Formeln für die Schätzfunktionen und Standardfehler lassen sich an komplexere Zufallsauswahlen

anpassen, bei denen nicht jedes Element der Grundgesamtheit die gleiche Chance hat, in die Stichprobe zu gelangen, und/oder mehrere Auswahlvorgänge hintereinander geschaltet werden. Voraussetzung ist aber, daß die Auswahlwahrscheinlichkeit bekannt ist und daß die einzelnen Ziehungen weiterhin unabhängig voneinander erfolgen (daß also die Auswahlchance eines Elements nicht davon abhängt, daß ein anderes Element gezogen wird). Die einführenden Standardtexte (wie z. B. Böltken 1976) bieten die entsprechenden Anpassungsformeln in der Regel nur für arithmetische Mittel und deren Standardfehler, nicht aber für multivariate Kennzahlen an. Hierzu muß man die einschlägige Spezialliteratur konsultieren. Die Anpassungsformeln werden oft so kompliziert, daß man in der Praxis kaum auf sie zurückgreift. Einer Daumenregel zufolge, soll man die Standardfehler mit $\sqrt{2}$ multiplizieren (ebd., S. 369).

9.3 Geschichtete Zufallsstichproben

Bei geschichteten Zufallsstichproben werden die Elemente der Grundgesamtheit anhand eines Merkmals X (oder mehrerer Merkmale) in Gruppen (»Schichten«, »strata«) eingeteilt, und zwar so, daß jedes Element zu einer - nur zu einer - Schicht gehört. So kann man z. B. eine Grundgesamtheit von Personen anhand des Religionsbekenntnisses in Protestanten, Katholiken, Juden und »Andere« einteilen. Aus diesen Schichten (Teilpopulationen) werden einfache Zufallsstichproben gezogen. Deren Umfänge kann man so festlegen, daß sie proportional zu den jeweiligen Anteilen der Schichten an der Grundgesamtheit variieren. Dann spricht man von »proportional geschichteten Stichproben«. Im anderen Falle bezeichnet man die Gesamtstichprobe als »disproportional geschichtete Stichprobe«.

Bei ihr haben die einzelnen Elemente der Grundgesamtheit also nicht alle dieselbe Auswahlchance. Bei der Datenauswertung müssen deshalb die Stichprobenelemente unterschiedlich gewichtet werden, wenn man Populationsparameter und Stadardabweichungen für die Schätzer ermitteln will. Die Gewichtung erfolgt mit dem reziproken Wert der Auswahlwahrscheinlichkeiten (man muß also die Verteilung des Schichtungsmerkmals in der Grundgesamtheit kennen). Datenanalysesysteme wie SPSS und BMDP stellen dafür eine spezielle Funktion WEIGHT bereit. Bei mehrstufigen Auswahlverfahren kann die Berechnung von Gewichtungsfaktoren eine ziemlich schwierige Angelegenheit werden.

Geschichtete Stichproben bieten vor allem zwei Vorteile:

(a) Man kann auf diese Weise sicherstellen, daß eine bestimmte Kategorie von Elementen (z. B. ethnische Minoritäten) in der Stichprobe in

ausreichender Zahl vertreten sind - »ausreichend« z. B. im Hinblick auf die Anwendbarkeit bestimmter statistischer Analysetechniken.

(b) Häufig können die Populationsparameter anhand der Statistiken aus geschichteten Zufallsstichproben bei gleichem Aufwand präziser geschätzt werden, als dies mit Ergebnissen aus einfachen Zufallsstichproben möglich wäre.

So ist z. B. der Standardfehler des arithmetischen Mittels $\bar{y}$ nur noch von den Varianzen s_y^2 **innerhalb** der einzelnen Schichten abhängig, nicht mehr von den Differenzen der Schichtmittelwerte zum Gesamtmittelwert wie im Falle der einfachen Zufallsstichprobe (vergl. das in Teil I Abschn. 4.2.5 erläuterte Konzept der Varianzaufteilung). Man bezeichnet diese Reduzierung des Standardfehlers als »Schichtungseffekt«. Er ist im allgemeinen um so größer, je homogener die Schichten und je unterschiedlicher ihre Mittelwerte sind. Verfügen z. B. alle Personen einer bestimmten Schicht über das gleiche Einkommen, so gäbe es bei wiederholten Stichprobenziehungen keinerlei Schwankungen im Durschnittseinkommen; der Standardfehler wäre auf Null reduziert. Allerdings können bei disproportionaler Schichtung u. U. auch Standardfehler auftreten, die größer sind als bei der einfachen Zufallsauswahl. Andererseits läßt sich mit einigem Geschick der Schichtungseffekt durch disproportionale Schichtung gegenüber der proportionalen noch vergrößern. Meistens ist der Forscher jedoch an mehreren Merkmalsdimensionen interessiert. Eine Schichteinteilung, die für ein bestimmtes Merkmal optimal sein mag, kann im Hinblick auf andere Merkmale eher ineffizient sein.

9.4 Klumpenstichproben

Bei den bisher besprochenen Verfahren wurden stets **einzelne** Elemente der Grundgesamtheit (mit oder ohne vorherige Schichtung) ausgewählt. Beim Klumpenverfahren (»cluster sampling«) wird das Prinzip der einfachen Zufallsauswahl auf zusammengefaßte Elemente angewandt. Innerhalb der (zufällig) ausgewählten Einheiten (»Klumpen«) werden dann (in der einfachen Version des Klumpenverfahrens) alle Elemente erhoben.

Dieses Verfahren bietet sich z. B. an, wenn keine »Liste« mit den Elementen der Grundgesamtheit vorhanden ist, wohl aber eine Liste aller »cluster«. Auch Kostenerwägungen können für dieses Verfahren sprechen. So mag es zu aufwendig sein, eine Liste aller Haushalte einer großen Stadt zusammenzustellen, während eine Liste von Wohnblocks mit Hilfe des Stadtbauamts relativ leicht zu bekommen ist. Anhand dieser Liste kann nun eine bestimmte Anzahl von Wohnblocks nach dem Zufallsprinzip ausgewählt werden. Anschließend versuchen Interviewer, alle Haushalte innerhalb der ausgewählten Wohnblocks aufzusuchen. Auf ähnliche Weise können bestimmte Aktenregale eines Archivs ausgewählt werden.

Beim Klumpenverfahren werden die Formeln für die Schätzfunktionen und Standardfehler ziemlich kompliziert, vor allem dann, wenn die Klumpen unterschiedlich große Mengen von Elementen enthalten. Zudem können Schätzungen, die auf Klumpenstichproben beruhen, im Vergleich zu Schätzungen, die auf einfachen Zufallsstichproben beruhen, recht ungenau sein (»Klumpeneffekt«). Diese Ungenauigkeit wird um so größer,

- je stärker sich die Elemente eines Clusters hinsichtlich der interessierenden Variablen ähneln (entsprechend schlecht bilden sie die Grundgesamtheit ab) und
- je stärker sich die Cluster hinsichtlich der untersuchten Merkmalsdimension unterscheiden
  (Diese beiden Effekte bedingen einander. Wenn die Klumpen völlig homogen sind, muß die Populationsvarianz gänzlich in Form von Unterschieden zwischen den Klumpen auftreten. Deshalb läßt sich z. B. der Standardfehler des arithmetischen Mittels allein aus der »between variance« der Klumpenmittelwerte errechnen, sofern die Klumpen vollständig erhoben worden sind. Bei der geschichteten Auswahl ist es gerade umgekehrt: es werden alle Schichten erhoben, aber aus den Schichten wird eine Zufallsauswahl getroffen; entscheidend ist also die »within-variance«.)
- je kleiner die Zahl der Klumpen und desto größer die Zahl der Elemente in ihnen
- je größer die Unterschiede in der Größe der Klumpen.

Positiv gewendet, lassen sich diese Tendenzen wie folgt zusammenfassen: Man ziehe möglichst viele, kleinere, gleich große Klumpen, die in sich möglichst heterogen und untereinander möglichst homogen sind (siehe Böltken 1976, S. 304). Im Gegensatz dazu gilt für die geschichtete Zufallsauswahl, daß die einzelnen Schichten untereinander möglichst heterogen und in sich möglichst homogen sein sollen. Außerdem sollen die Schichtungsmerkmale, nicht aber die Klumpenmerkmale mit den Erhebungsmerkmalen korrelieren (ebd., S. 308).

Falls die Klumpengröße nicht konstant gehalten werden kann, empfiehlt sich eine größenproportionale Auswahl. Werden bei einer Klumpenstichprobe nicht alle Elemente der ausgewählten Cluster erhoben, sondern wird ihnen jeweils eine einfache Zufallsstichprobe entnommen, spricht man von einer zweistufigen Auswahl.

9.5 Mehrstufige Zufallsauswahlen

Bei ihnen wird die Grundgesamtheit zunächst in Gruppen von Elementen eingeteilt (z. B. im Sinne der »Klumpen« des vorigen Abschnitts). Man bezeichnet sie als Primäreinheiten (»primary sampling units« - PSU). Aus ihnen wird eine Zufallsauswahl gezogen, aus denen in einer zweiten Auswahlstufe eine Zufallsstichprobe der Sekundäreinheiten entnommen wird. Diese Sekundäreinheiten können bereits die Erhebungseinheiten bilden oder als Grundlage einer weiteren Auswahlstufe dienen usw. »Mehrstufige Auswahlverfahren bestehen also aus einer Reihe nacheinander durchgeführter Zufallsstichproben, wobei die jeweils entstehende Zufallsstichprobe die Auswahlgrundlage der folgenden Zufallsstichprobe darstellt« (Schnell et al. 1988, S. 265). Mehrstufige Auswahlverfahren werden z. B. angewandt, wenn eine »repräsentative« Bevölkerungsstichprobe erhoben werden soll[2].

In der Regel bilden räumliche Einheiten (wie Stimmbezirke, Kreise) die Primäreinheiten, aus denen in der ersten Stufe eine Zufallsstichprobe gezogen wird. Daraus direkt oder über weitere Auswahlstufen, können Haushalts- und Personenstichproben gezogen werden.

Fast immer enthalten die Primäreinheiten unterschiedlich viele Elemente: die Stimmbezirke z. B. unterschiedlich viele Wahlberechtigte. Um sicherzustellen, daß dennoch alle Elemente im Endeffekt die gleiche Auswahlwahrscheinlichkeit haben, kann man z. B. wie folgt verfahren: Auf der ersten Stufe gibt man Auswahlwahrscheinlichkeiten vor, die proportional zur Größe der Primäreinheiten sind. Auf der nächsten Stufe zieht man sodann dieselbe Zahl von Sekundäreinheiten. Diesen Ansatz bezeichnet man als PPS-Design (»probability proportional to size«).

Ein kompliziertes mehrstufiges Auswahlverfahren für Bevölkerungsumfragen hat z. B. der Arbeitskreis deutscher Marktforschungsinstitute (ADM) entwickelt. Darüber berichtet Kirschner (1984).

---

[2] Der Begriff »repräsentative Stichprobe« wird gelegentlich als Synonym für »Zufallsstichprobe« gebraucht, häufig aber auch auf andere Auswahlverfahren (z. B. Quotaverfahren, siehe unten) bezogen, die nicht den Prinzipien der Zufallsauswahl folgen. Im ersten Fall ist der Begriff der Repräsentativität überflüssig, im zweiten meist nicht eindeutig definiert oder gar irreführend. Für die Inferenzstatistik ist der Begriff jedenfalls irrelevant. Man kann jedoch die »Repräsentativität« der Stichprobe, d. h. die Übereinstimmung von Merkmalsverteilungen der Stichprobe mit den entsprechenden Verteilungen der Grundgesamtheit, als Indikator dafür nehmen, daß eine vorliegende Stichprobe einer Zufallsstichprobe hinsichtlich diser Merkmale ähnlich ist.

9.6 Das Quotaverfahren

Stichproben, die nicht nach dem Zufallsprinzip, aber trotzdem nach festgelegten Regeln gezogen werden, bezeichnet man als **bewußte** Auswahlen. Die wohl bekannteste Form einer bewußten, nicht-zufälligen Auswahl ist das Quotaverfahren. Es wird vor allem in der Demoskopie und in der Marktforschung angewandt. Dieses Verfahren beruht darauf, daß man die Verteilung bestimmter Merkmale (der »Quotenmerkmale«) in der Grundgesamtheit kennt. So weiß man z. B. aus Volkszählungen und sonstigen amtlichen Statistiken, wie groß die Anteile der Männer und der Frauen oder der über 60jährigen und der Achtzehn- bis Dreißigjährigen in der Bundesrepublik sind. Man versucht nun, die Stichprobenziehung so zu steuern, daß die Quotenmerkmale in der Stichprobe möglichst genau so verteilt sind wie in der Grundgesamtheit; daß also z. B. die relative Häufigkeit der Männer in der Stichprobe so groß ist wie in der Population. Statt eine solche »Repräsentativität« über eine Zufallsauswahl sicherzustellen, überläßt man den einzelnen Interviewern die Auswahl der zu befragenden Personen. Man gibt ihnen lediglich vor, wieviele Personen eines bestimmten Merkmals sie jeweils befragen sollen. Diese »Quotenvorgaben« beruhen auf dem »Quotenplan«, der so berechnet ist, daß bei korrekter Arbeit der Interviewer die Verteilung der quotierten Merkmale in der Stichprobe exakt ihrer Verteilung in der Grundgesamtheit entspricht. Es können mehrere Merkmale unabhängig voneinander oder in kombinierter Form quotiert werden. Im letzten Falle kommt es z. B. nicht nur darauf an, daß ein bestimmter Anteil p(F) an Frauen und ein bestimmter Anteil p (A) an Achtzehn- bis Dreißigjährigen interviewt wird, sondern ein bestimmter Anteil p (F,A) an achtzehn- bis dreißigjährigen Frauen.

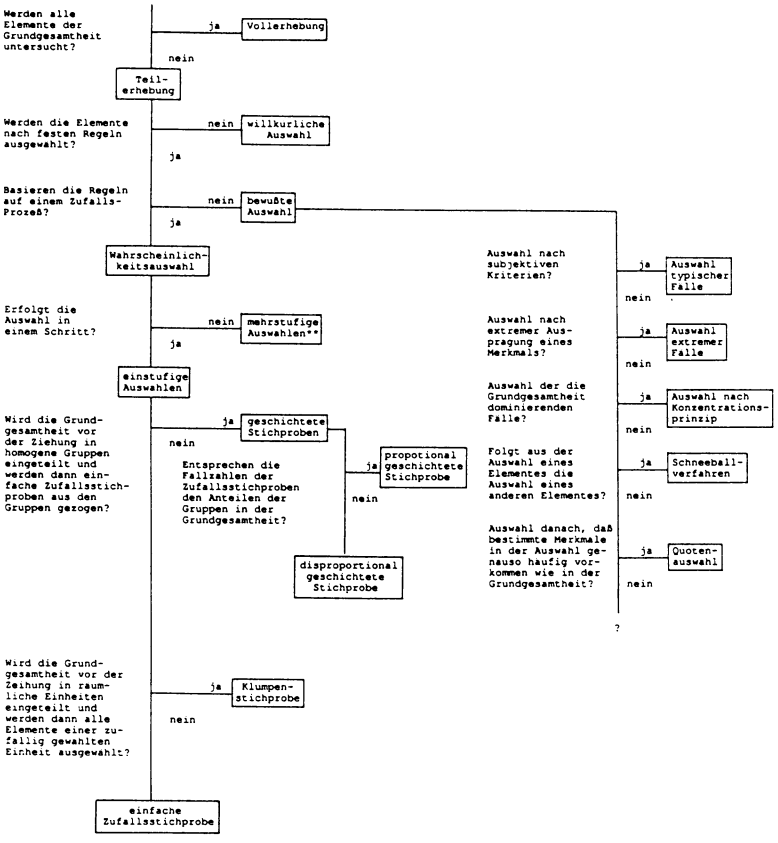
Das Quotenverfahren wird vor allem aus praktischen und finanziellen Gründen gewählt; es ist in der Regel wesentlich billiger als die Zufallsauswahl. Dafür nimmt man gravierende Nachteile in Kauf:

(a) Es entfällt die Grundlage für die Anwendung der Inferenzstatistik.
(b) Es besteht die Gefahr erheblicher Stichprobenverzerrungen
   - durch die höhere Auswahlwahrscheinlichkeit für Personen, die häufig zu Hause sind,
   - durch die Interviewer, die vor allem Personen auswählen, die für sie leicht zugänglich sind (Freunde, Bekannte) und die sich kooperativ verhalten.

Man versucht, diese Gefahren durch Schulung und Kontrolle der Interviewer zu mindern, völlig ausschalten lassen sie sich nicht. Schnell et al. (1988, S. 279) resümieren: »Falls exakte Ergebnisse, deren Genauigkeit angebbar ist, wichtiger sind als 'kostengünstige' Lösungen, gibt es im allgemeinen keine Alternative zu Zufallsstichproben.«

Zum Schluß reproduzieren wir aus dieser Arbeit noch eine Übersicht, in der die verschiedenen Auswahlverfahren (die wir hier weder ausführlich noch vollständig besprochen haben) mit ihren charakteristischen Unterschieden einander gegenübergestellt werden.

Abb. 9.1: Auswahlverfahren



** Mehrstufige Auswahlen bestehen aus Kombination einstufiger Verfahren mit unterschiedlichen Auswahleinheiten

Quelle: Schnell/Hill/Esser 1988, S. 255 (modifiziert)

KAPITEL 10

Bivariate Verteilungen II: Einfache Regressionsanalyse

10.1 Deskription: Bestimmung der Regressionsgeraden

Wir verlassen vorübergehend die inferenzstatistische Perspektive und kehren zur rein deskriptiven Betrachtung bivariater Verteilungen zurück. Dabei geht es, wie schon in Teil I, Abschnitt 4.2.4, um den Zusammenhang zweier metrischer Variablen, deren bivariate Verteilung mit zusätzlichen Kennzahlen, den Regressionskoeffizienten und den Determinationskoeffizienten, charakterisiert werden soll.

Um den Grundgedanken der Regressionsanalyse zu verdeutlichen, betrachten wir zunächst eine fiktive experimentelle Situation außerhalb der Historischen Sozialforschung. Nehmen wir an, ein Agrarwissenschaftler wolle den jährlichen Hektarertrag eines Getreides (Variable Y) in Abhängigkeit von unterschiedlichen Mengen eines eingesetzten Düngemittels (Variable X) untersuchen. Die unterschiedlich festgelegten Experimentiermengen x_{ik} des Düngemittels (gemessen in kg/ha) bringt er auf mehreren Versuchsfeldern aus und registriert später die Erträge y_{ik} ($k = 1,2, \ldots, K$) indiziert die K unterschiedlichen Mengen des eingesetzten Düngemittels; »i«, $i = 1,2, \ldots, n$ die durchnumerierten Versuchsfelder). Er achtet darauf, daß die Bodenqualität möglichst konstant ist. Abb. 10.1 zeigt ein denkbares Ergebnis des Experiments.

Bei jeder gegebenen Menge x_k des Düngemittels werden unterschiedliche Ertragsmengen y_{ik} beobachtet. Sie streuen um einen bedingten Mittelwert $\bar{y}_k \mid x_k$. Diese arithmetischen Mittel variieren systematisch mit der Menge x_k; der Zusammenhang ist aber nicht linear. Abweichungen von den durchschnittlichen Erträgen können durch Meßfehler und unkontrollierbare Einflußfaktoren (nicht völlig konstante Bodenqualität, nicht völlig konstante Niederschlagsmenge und Sonnenbestrahlung) entstehen. Da $\bar{y}_k \mid x_k$ das arithmetische Mittel aller unter der Bedingung x_k erzielten Erträge darstellt, gleichen sich die positiven und negativen Abweichungen e_{ik} in der Summe aus: $E(e_{ik}) = 0$. Bei einer Wiederholung des Experiments (bzw. bei einer Prognose) lassen sich die arithmetischen Mittel $\bar{y}_k \mid x_k = E(Y \mid x_k)$ als bedingte Erwartungswerte interpretieren. Die Linie, die die einzelnen Erwartungswerte (arithmetischen Mittel) miteinander verbindet, wird als **Regressionslinie** bezeichnet.

Dieses aus der experimentellen Wissenschaft stammende Regressionsmodell läßt sich für die nicht-experimentellen Wissenschaften modifizieren, wenn man

(a) auch die unabhängige Variable X und nicht nur die unabhängige Variable Y als **Zufalls**variable zuläßt,
(b) die Regressionslinie und damit die Erwartungswerte $E(Y \mid x_i)$ für jede Realisierung $X = x_i$ aus den Beobachtungsdaten aller Y- und X-Werte mathematisch konstruiert. (Da die X-Werte jetzt nicht mehr experimentell festgelegt, sondern für jede Untersuchungseinheit gesondert erhoben werden, entfällt der zweite Index, k.)

Dabei setzt man häufig vereinfachend voraus, daß der Zusammenhang zwischen der Variable Y (die man auch als Kriteriumsvariable bezeichnet) und der Variable X (die man auch als Prädiktor- oder Regressorvariable bezeichnet) **linear** sei. Unter Umständen müssen die Variablen zuvor transformiert (z. B. logarithmiert) werden, um diese Voraussetzung zu erfüllen (siehe Kap. 12.1). Gelegentlich gibt es jedoch Variablenbeziehungen, die sich nicht durch Transformation »linearisieren« lassen (siehe Kap. 12.2). Auf sie ist das jetzt zu besprechende Regressionsverfahren (»lineare« Regression) nicht anwendbar.

Für unser sozialwissenschaftliches Beispiel zur Regressionsanalyse greifen wir auf das Streudiagramm zurück, das uns schon in Teil I, Kap. 4.2.4 bei der Ableitung des Pearsonschen Produkt-Moment-Korrelationskoeffizienten r als Vorlage diente, hier reproduziert als Abb. 10.2. Es zeigt die bivariate Verteilung der Variable Y:= Stimmenanteile der SPD in den Wahlbezirken bei der Wahl zum Reichstag 1912 und der Variable X:= Anteil der Wahlberechtigten, die in Industrie oder Gewerbe beschäftigt sind (nebst den Haushaltsangehörigen).

In der beschreibenden Analyse versuchen wir, die Vielfalt der Informationen zu reduzieren, das Wesentliche einer uni- oder multivariaten Verteilung in wenigen statistischen Kennzahlen auszudrücken. Bei zwei- und mehrvariaten Verteilungen gehen wir von bestimmten Annahmen (einem »Modell«) über Form und Struktur der Beziehung zwischen den Variablen aus (siehe Abschn. 10.2). So setzt man häufig voraus, daß die Beziehung monoton oder linear ist (oder durch entsprechende Transformationen linear »gemacht« werden kann). Ein Blick auf unser Streudiagramm (Abb. 10.2) läßt diese Annahme als vertretbar erscheinen. Es vermittelt den Eindruck, daß die Stimmenanteile der SPD in den einzelnen Wahlbezirken der Tendenz nach proportional zur jeweils erreichten industriellen Entwicklung zunehmen. Allerdings könnte sich die Steigung nach rechts ein wenig abflachen. Wir wollen aber zunächst mit dem Modell einer linearen Beziehung arbeiten[1].

[1] Die Linearitätsannahme und andere Modellvoraussetzungen (siehe

Die Annahme einer linearen Beziehung läßt sich durch eine einfache Gleichung formalisieren:

$$(10\text{-}1) \qquad \hat{Y} = a + bX$$

Wir haben Y mit einem Dach versehen, weil wir zunächst nur eine »Tendenzaussage« machen wollen über »durchschnittliche« Y-Werte bei unterschiedlichen X-Werten. Diese Tendenz läßt sich durch eine Gerade ausdrücken, die wir nach bestimmten Kriterien durch den »Punktschwarm« des Streudiagramms (siehe Abb. 10.2) legen. Die »Regressionsgerade« soll den Punktschwarm möglichst gut repräsentieren (was das heißt, wird sogleich deutlich werden). Sie ist eindeutig bestimmt, wenn gemäß Gleichung (10-1) die sog. Regressionskoeffizienten, a und b, festgelegt sind. Schematisch wird das in Abb. 10.3 verdeutlicht.

Man bezeichnet »a« als Ordinatenabschnitt (»intercept«), als Regressionskonstante oder Absolutglied und »b« als **Steigungskoeffizienten** (»slope«) oder Regressionsgewicht. Der Ausdruck »Steigungskoeffizient« verdankt sich dem Tatbestand, daß b gleich dem Tangens des Winkels α ist, den die Gerade mit einer Abszissenparallele bildet. Für zwei beliebige Punkte $(x_1; \hat{y}_1)$ und $(x_2; \hat{y}_2)$ auf der Geraden gilt:

$$(10\text{-}2) \qquad \hat{y}_1 = a + bx_1$$

$$\hat{y}_2 = a + bx_2$$

$$\hat{y}_1 - \hat{y}_2 = b(x_1 - x_2)$$

$$\frac{\hat{y}_1 - \hat{y}_2}{x_1 - x_2} = b$$

Die Differenzen $(\hat{y}_1 - \hat{y}_2)$ und $(x_1 - x_2)$ bilden Gegenkathete und Ankathete in einem von der Regressionsgeraden als Hypothenuse vervollständigten

---

Abschn. 10.2) sind in diesem Beispiel aus theoretischen Gründen nicht unproblematisch, da die Prozentangaben für die Stimmenanteile der SPD nach unten (0%) und oben (100%) begrenzt sind. In der sozialwissenschaftlichen Forschungspraxis bleiben daraus resultierende Probleme meist unbeachtet. Wir wollen sie auch hier vorläufig übergehen, aber in Kap. 12 ausführlich behandeln. - Ein weiterer Problempunkt unseres Analysebeispiels liegt darin, daß nicht Individuen, sondern Kollektive (Wahlkreise) die Untersuchungseinheiten darstellen. Zwar ist die einzelne Stimmabgabe für eine bestimmte Partei Merkmal einer Person; aber die Stimmenanteile für die SPD (oder eine andere Partei) sind (analytische) Kollektivmerkmale der Wahlkreise. Das muß bei der Interpretation der Ergebnisse berücksichtigt werden. Darauf werden wir ebenfalls in Kap. 12 zurückkommen.

Dreieck. Das Verhältnis von Gegenkathete zu Ankathete in einem rechtwinkligen Dreieck bezeichnet man als »Tangens« des Winkels α, der durch Hypothenuse und Ankathete eingeschlossen wird. Der Steigungskoeffizient b gibt also an, um wieviel Einheiten sich Y im Durchschnitt verändert, wenn X sich um eine Einheit ändert[2]. Wenn $x_1 - x_2 = 1$, wird die letzte Zeile in Gleichung (10-2) zu $\hat{y}_1 - \hat{y}_2 = b$. Der Ordinatenabschnitt a ist häufig nicht inhaltlich interpretierbar.

Wie wird nun die Gerade, d. h., wie werden die Regressionskoeffizienten a und b bestimmt? Ein naheliegendes Kriterium ist die Forderung, die Gerade so zu legen, daß die einzelnen Punkte möglichst gering von ihr abweichen. Um diese Forderung in ein formales Kriterium zu übersetzen, um die Aufgabe also rechnerisch lösbar zu machen, müssen wir Gleichung (10-1) erweitern, indem wir eine Fehlergröße, e_i, mit berücksichtigen:

$$
\begin{aligned}
(10\text{-}3) \qquad & y_1 = a + bx_1 + e_1 \\
& \vdots \qquad\quad \vdots \\
& y_n = a + bx_n + e_n
\end{aligned}
$$

Der Index i bezeichnet wiederum die durchnumerierten Untersuchungseinheiten. Für jede von ihnen läßt sich der beobachtete Wert $Y = y_i$ als eine Funktion a) der X-Variablen und b) einer »Fehler«- oder »Residualvariable« (»error«, e) schreiben. Wir erhalten also nicht nur eine Gleichung, sondern ein System von Gleichungen. Der Begriff des »Fehlers« (man spricht auch von »Störgrößen«) folgt auch in diesem Falle (wie schon bei den in Teil I besprochenen PRE-Maßen) aus einer Prognosekonzeption: Wenn wir die Y-Werte mit Hilfe der X-Werte und der Regressionskoeffizienten prognostizieren, machen wir Fehler, wenn Y nicht vollständig durch X determiniert ist. Auf diese Prognosekonzeption werden wir gleich zurückkommen. Wenn wir nun fordern, die Gerade so zu legen, daß der Fehler möglichst gering wird, läßt sich diese Forderung formal in Gleichung (10-4) ausdrücken:

$$(10\text{-}4) \qquad \sum |e_i| = \sum |y_i - (a + bx_i)| = \min.$$

---

[2] Formal erhält man die Steigung einer Kurve oder Geraden über die 1. Ableitung der entsprechenden Funktion, hier Gleichung (10-1): $\delta\hat{y}/\delta x = b$. In dem Falle einer Geraden ist die Steigung eine Konstante.

Aus Gründen, die wir nicht erörtern wollen, wird aber in der Regel ein etwas anderes Kriterium vorgezogen:

$$(10\text{-}5) \qquad \sum_{i=1}^{n} e_i^2 = \min.$$

Die Koeffizienten a und b sollen so gewählt werden, daß die Summe der Fehlerquadrate ein Minimum darstellt (Methode der kleinsten Fehlerquadrate - KFQ; »Ordinary Least Squares« - OLS). Wir interpretieren folglich Σe_i^2 als ein Funktion der zu bestimmenden Regressionskoeffizienten bei gegebenen (nämlich beobachteten) X- und Y-Werten. Wir betrachten vorübergehend nicht X und Y, sondern die Regressionskoeffizienten als Variablen. Die Fehlerquadratsumme wird bei extrem niedrigen oder extrem hohen Regressionskoeffizienten groß und irgendwo dazwischen minimal. Dieses Minimum müssen wir bestimmen (bei welchem a und welchem b wird es erreicht?). Abbildung 10.4 zeigt das im Hinblick auf den Steigungskoeffizienten b bei konstantgehaltenem Ordinatenabschnitt a. In entsprechender Weise können wir auch a variieren und b konstanthalten. Aus der Schulmathematik ist bekannt, daß man Maxima und Minima allgemein mit Hilfe der ersten beiden Ableitungen bestimmt. Hier benötigen wir nur die 1. Ableitungen:

$$(10\text{-}6) \qquad \frac{\partial[\sum_{i=1}^{n} (y_i - a - bx_i)^2]}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial[\sum_{i=1}^{n} (y_i - a - bx_i)^2]}{\partial b} = -2 \sum_{i=1}^{n} x_i(y_i - a - bx_i)$$

Setzt man diese partiellen Ableitungen gleich Null führen sie zu den sog. »Normalgleichungen«

$$(10\text{-}7) \qquad n \cdot a + \sum x_i b = \sum y_i$$

$$\sum x_i \cdot a + \sum x_i^2 \cdot b = \sum x_i y_i$$

Deren Lösungen für a und b lauten:

$$(10\text{-}8) \qquad a = \bar{y} - b\bar{x}$$

$$b = \frac{1/n \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{1/n \sum(x_i - \bar{x})^2}$$

Diese Informationen können wir mit SPSS[x] über verschiedene Kommandos abrufen; die Standardprozedur (in ihrer einfachsten Ausführung) ist:

```
REGRESSION VARIABLES=PROZSPD INDUSTRY
/DEPENDENT=PROZSPD/ENTER
```

Die Prozedur kann durch eine Reihe von Subkommandos ergänzt werden, die weitere Informationen liefern, von denen wir noch einige kennenlernen werden.

In unserem Beispiel erhalten wir die folgenden (empirischen) Regressionskoeffizienten:

$$(10\text{-}9) \qquad a = -6{,}78$$

$$b = 0{,}80$$

Hier wird deutlich, daß der Ordinatenabschnitt offensichtlich durch die univariaten Verteilungen in X und Y bestimmt ist (der Wert 0 wird in Y beobachtet, aber nicht in X) und somit über die Beziehung der beiden Variablen nichts aussagt. Der unrealistische negative Betrag ist aber in diesem Falle ein Hinweis darauf, daß die Beziehung nicht korrekt spezifiziert ist (siehe obige Fn. 1). Der Steigungskoeffizient besagt, daß die SPD im Schnitt pro Prozentpunkt Beschäftigtenzuwachs in der Industrie 0,80 Prozentpunkte an Stimmen gewonnen hat. Für einen Wahlkreis, in dem 40 % der Bevölkerung in Industrie und Gewerbe beschäftigt sind, würde man also laut Modell einen SPD-Stimmenanteil von $\hat{y} = -6{,}78\,\% + 0{,}80 \cdot 40\,\% = 25{,}22\,\%$ erwarten. Leider liegen keine fortlaufenden Erhebungen zum Industrialisierungsgrad vor, so daß wir die Stabilität dieser Beziehung über Zeit (im Vergleich mehrerer Reichstagswahlen) nicht untersuchen können.

Die Bestimmungsgleichung (10-8) für a läßt erkennen (wenn man sie nach $\bar{y}$ umstellt), daß die nach der Kleinstquadratmethode ermittelte Gerade stets durch den Punkt $(\bar{x};\bar{y})$ verläuft. Man bezeichnet ihn als den »Schwerpunkt« der bivariaten Verteilung von X und Y.

Die Bestimmungsgleichung für den Steigungskoeffizienten b enthält im Zähler den gleichen Ausdruck wie Pearsons Produkt-Moment-Korrelationskoeffizient, nämlich die Kovarianz (siehe die Gleichungen (4-31) und (4-34) in Teil I).

$$(10\text{-}10) \qquad r = \frac{1/n\Sigma(\bar{x}_i - x)(y_i - \bar{y})}{\sqrt{1/n(x_i-\bar{x})(y_i-\bar{y})^2}} = \frac{Kov(x,y)}{s_x s_y}$$

Dividiert wird aber nicht (wie bei r) durch die beiden Standardabweichungen s_x und s_y, sondern lediglich durch die Varianz der X-Variablen. Damit wird b zu einem asymmetrischen Maß. Wenn die Varianzen der beiden Variablen nicht identisch sind, erhält man bei der Regression von Y auf X einen anderen Steigungskoeffizienten, b_{yx}, als bei der Regression von X auf Y, b_{xy}. Um eventuelle Zweifel auszuschließen, auf welches Kriterium (abhängige Variable) und welchen Regressor (unabhängige Variable) sich der Steigungskoeffizient bezieht, versieht man ihn gelegentlich mit einem doppelten Index, in dem die Kriteriumsvariable zuerst genannt wird. Somit gelten die Beziehungen

$$(10\text{-}11) \qquad b_{yx} = r(s_y/s_y) \qquad r = b_{yx}(s_x/s_y)$$

Nachdem wir die »optimale« Regressionsgerade bestimmt haben, benötigen wir noch eine Kennzahl, die angibt, wie »gut« die Gerade die Punktwolke repräsentiert, in welchem Umfang die beobachteten Werte um sie streuen. Es liegt nahe, hierzu den mittleren quadratischen Fehler (MQF oder MSE = Mean Square Error) zu verwenden:

$$(10\text{-}12) \qquad MQF = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Wenn man statt durch die Zahl n der Fälle durch die Freiheitsgrade df = n - 2 (wenn zwei Regressionsparameter, a und b, bestimmt werden, können nur n - 2 Fälle unabhängig voneinander variieren) dividiert, erhält man die sog. **Residualvarianz** s_e^2. Bisher (in Teil I) waren Streuung bzw. Varianz definiert im Hinblick auf die Abweichung der einzelnen Werte vom arithmetischen Mittel der Variable. Jetzt definieren wir sie im Hinblick auf die Abweichung der einzelnen Werte y_i ($i = 1,2, \ldots, n$) von den

entsprechenden Punkten der Regressionsgeraden, den bedingten Erwartungswerten $E(Y \mid x_i) = \hat{y}_i$. Die Regressionsgerade war ja unter der Bedingung gefunden worden, die Summe der quadrierten Differenzen $[\, y_i - (a + bx_i)\,]$ zu minimieren. Das bietet uns den Ansatzpunkt, nicht bei dem mittleren quadratischen Fehler als Gütemaß für die Regressionsgerade stehenzubleiben, sondern ein PRE-Maß (siehe Teil I, Kap. 4) zu entwickeln. (Der MQF ist offensichtlich abhängig von der skalenbedingten Größe der Y-Werte.)

Dazu fingieren wir wieder eine Prognosesituation (vergl. Teil I, Kap. 4), in der wir die einzelnen y_i-Werte auf zweierlei Weise voraussagen: a) ohne Kenntnis der x_i-Werte, b) auf der Basis der x_i-Werte. Bei der Prognose ohne Kenntnis der x_i-Werte minimieren wir die Summe unserer Fehlerquadrate, indem wir jedesmal das arithmetische Mittel $\bar{y}$ als Schätzer für y_i ($i = 1,2, \ldots, n$) verwenden. Falls wir die x_i-Werte kennen (und die vorausgesetzte Linearität der Beziehung zutrifft) ist unsere beste Prognose (im Sinne des mittleren quadratischen Fehlers) der jeweilige Punkt $\hat{y}_i$ auf der Regressionsgeraden; denn unter dieser Voraussetzung wurde sie ja abgeleitet (siehe nochmals Abb. 10.3). Wir haben also wiederum zwei Fehlertypen, E_1 und E_2, die wir ins Verhältnis zueinander setzen können:

$$(10\text{-}13) \qquad PRE = \frac{E_1 - E_2}{E_1} = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

Analog zur Ableitung des Koeffizienten Eta^2 (siehe Teil I, Abschn. 4.2.5.) können wir auch dieses PRE-Maß, das man als **Determinationskoeffizienten** bezeichnet, als Resultat einer Zerlegung der Variation $\Sigma(y_i - \bar{y})^2$, der Summe der Abweichungsquadrate, betrachten. Statt der Gruppenmittelwerte $\bar{y}_j$ in den Gleichungen (4-40 ff.) setzen wir hier die $\hat{y}_i$ als Prognosewerte ein. Die Zerlegung sieht somit wie folgt aus: Die Differenz $(y_i - \bar{y})$ kann in zwei Komponenten, zwei Streckenabschnitte, eingeteilt werden:

$$(10\text{-}14) \qquad (y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Durch Quadrieren, Ausmultiplizieren des Binoms (der mittlere Ausdruck des Binoms wird Null[3]) und Summieren erhält man daraus folgende Gleichung:

---

[3] Voraussetzung hierfür ist, daß die Störgröße e_i und die Prognosewerte $\hat{y}_i$ nicht miteinander korrelieren (siehe Pyndick/Rubinfeld 1981, S. 79).

$$(10\text{-}15) \quad \Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$$

$$\text{Gesamtvariation} = \text{erklärte Variation} + \text{nichterkl. Variation}$$

Durch Einsetzen von (10-15) in (10-13) erhalten wir

$$(10\text{-}16) \quad PRE_D = \frac{\sum(y_i-\bar{y})^2 - \sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

$$= \frac{[\Sigma(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2] - \sum(y_i-\hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$$= \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

$$PRE_D = \frac{\text{erklärte Variation}}{\text{Gesamtvariation}}$$

Es läßt sich zeigen (siehe Schlittgen 1987, 369 f.), daß dieser Determinationskoeffizient gleich dem quadrierten Pearsonschen Korrelationskoeffizienten r ist. Man verwendet deshalb für ihn im allgemeinen das Symbol r^2 oder R^2. In unserem Analysebeispiel ist $R^2 = 0{,}50$. Demnach werden unter der Voraussetzung eines linearen Modells 50 % der Varianz der SPD-Stimmenanteile allein durch den Industrialisierungsgrad der Wahlkreise »erklärt«. Komplementär zum Determinationskoeffizienten (auch als »Bestimmtheitsmaß« bezeichnet) ist ein Unbestimmtheitsmaß (»coefficient of alienation«) definiert: $1 - R^2$. Es gibt an, wie groß der Anteil der nicht-erklärten Variation an der Gesamtvariation ist.

Damit haben wir den etwas vagen Begriff des (linearen) »Zusammenhangs« zwischen zwei Variablen nach zwei Aspekten differenziert (vergl. Schluß des Abschnitts 4.2.5 in Teil I):

(1) in ein Maß dafür, wie stark sich Y pro Einheitsänderung in X verändert (Steigungskoeffizient b)
(2) in ein Maß dafür, wie groß die Abweichungen von dieser durchschnittlichen Änderung sind, wie stark die Variation der Y-Werte durch die Variation der X-Werte »gebunden« ist (Determinationskoeffizient R^2).

In der Kausalanalyse wird in der Regel der Steigungskoeffizient, den man als »Strukturparameter« bezeichnet, als aussagekräftiger angesehen.

Eine wichtige Eigenschaft, in der sich Steigungskoeffizient und Determinationskoeffizient unterscheiden, beleuchtet Abbildung 10.5.

In ihr werden zwei Situationen voneinander unterschieden. In der ersten liegen Daten vor, die die volle Spannweite der X-Werte darstellen. In der zweiten liegen zu den gleichen Varialen nur die eingerahmten Daten mit einer deutlich geringeren Spannweite der X-Variablen vor, während die Variation der Y-Werte um die Regressionsgerade konstant ist. In beiden Situationen hat die Regressionsgerade dieselbe Steigung. Im ersten Fall ist der Anteil der nicht-erklärten Varianz an der Gesamtvarianz offenkundig viel geringer als im zweiten Fall. Bei gleichem Steigungskoeffizienten ergeben sich also unterschiedlich große Determinationskoeffizienten. Allgemein gilt: bei gegebener Steigung der Geraden und gegebener Variation der Y-Werte um die Regressionsgerade ist der Determinationskoeffizient um so geringer, je geringer die Varianz in X; offensichtlich wird das **Verhältnis** der nicht-erklärten zur Gesamtvarianz um so größer, je kleiner die Spannbreite der beobachteten X-Werte ist. Hier haben wir eine Entsprechung zur Abhängigkeit vieler Koeffizienten der Tabellenanalyse von den univariaten Randverteilungen der jeweiligen Variablen. Das ist vor allem zu beachten, wenn Korrelationsergebnisse miteinander verglichen werden, die aus unterschiedlichen Populationen bzw. Stichproben stammen, in denen die Varianzen der X-Variable stark differieren.

Der Steigungskoeffizient hingegen ist nicht von der Varianz der X-Werte abhängig. Wenn die X-Werte gegeben sind, ist er als Linearkombination der Y-Werte darstellbar (siehe z. B. Schlittgen 1987, S. 375). Allerdings ist der Steigungskoeffizient von Stichprobe zu Stichprobe um so weniger stabil, sein Standardfehler ist um so größer, je geringer die Varianz in X (näheres hierzu siehe Abschnitt 10.3). Wie Abbildung 10.5 verdeutlicht, müssen sich Zufallseinflüsse, die bei der Stichprobenziehung auf Y einwirken, um so stärker auf die Steigung der Geraden auswirken, um so geringer die erfaßte Spannweite der X-Variable ist. Daraus ergibt sich die Forderung an den Empiriker: Maximiere die Varianzen der unabhängigen Variablen.

In unserem Beispiel müssen wir noch beachten, daß die Wahlkreise nicht die jeweils gleiche Zahl von Wahlberechtigten repräsentieren. Dadurch können sowohl der Ordinatenabschnitt als auch der Steigungskoeffizient beeinflußt sein, falls die Größe der Wahlkreise mit dem Industrialisierungsgrad und dem SPD-Stimmenanteil korreliert. Das ist in unserem Beispiel der Fall: Die Zahl der Wahlberechtigten korreliert sowohl mit dem Industrialisierungsgrad ($r = 0{,}44$) als auch mit dem SPD-Stimmenan-

teil (r = 0,43). Wenn es darum geht, die Bedeutung der Industrialisierung für den Stimmenanteil der SPD richtig einzuschätzen, möchte man den Einfluß ungleicher Wahlkreisgrößen ausschalten. Das läßt sich bewerkstelligen, indem man die Wahlkreise (also die einzelnen Fälle) proportional zur Zahl der Wahlberechtigten (registriert in der Variablen BERECH12) gewichtet. Das bedeutet, daß man einen Wahlkreis A praktisch zweimal zählt, wenn er doppelt so groß ist wie der Durchschnittswahlkreis.

Die Gewichtungsvariable wird in SPSS^x durch einen COMPUTE-Befehl gebildet und durch das WEIGHT-Kommando ausgeführt:

```
COMPUTE GEWICHT=(BERECH12/14366567)*395
WEIGHT BY GEWICHT
```

Diese beiden Befehle müssen vor das Prozedurkommando REGRESSION gesetzt werden. Es empfiehlt sich, nicht einfach mit der Zahl der Wahlberechtigten in dem jeweiligen Wahlkreis (BERECH12) zu gewichten, sondern eine Gewichtung zu wählen, die die Gesamtzahl der Fälle unverändert läßt. Das wird erreicht, indem man zunächst die Zahl der Wahlberechtigten des Wahlkreises ins Verhältnis setzt zur Zahl der Wahlberechtigten in allen 395 Wahlkreisen und dann diesen Ausdruck mit der Zahl der Wahlkreise multipliziert (für zwei Wahlkreise liegen keine Angaben vor). Würde man jeden Wahlkreis nur mit der Zahl seiner Wahlberechtigten gewichten, ergäbe sich eine rechnerische Fallzahl von über 14 Millionen, was Signifikanztests (siehe Abschnitt 10.3) wertlos machen würde[4]. Die Regressionskoeffizienten selbst werden aber nicht anders geschätzt, wenn man die Gewichtung mit WEIGHT BY BERECH12, also ohne »Normierung« auf 395 Fälle, eingibt, da die Proportionalitäten die gleichen bleiben.

Die Gewichtung führt dazu, daß wir mit dem neuen Regressionsmodell diejenigen Stimmenanteile schätzen, die die SPD in Abhängigkeit vom Industrialisierungsgrad erhalten hätte, wenn, unter sonst gleichbleibenden Voraussetzungen, die Größe der Wahlkreise konstant gewesen wäre. Da die Zahl der Wahlberechtigten in den einzelnen Wahlkreisen positiv mit den SPD-Stimmenanteilen korreliert, erhöht sich deren Mittelwert von knapp 24 auf etwas über 29 %. Der Niveauanstieg zeigt sich auch in dem Ordinatenabschnitt, der von a = - 6,78 auf a = - 1,42 zunimmt. Allerdings

---

[4] Da wir es hier, abgesehen von den fehlenden Werten zweier Wahlkreise, mit der Gesamtheit der Wahlkreise zu tun haben, erübrigen sich Signifikanztests, sofern man nicht das Konzept der theoretischen Population anwenden will, das alle empirischen Daten als Stichprobenrealisierungen betrachtet (siehe die einleitenden Bemerkungen zu Kap. 6). Aber auch unabhängig von dieser Konzeption wollen wir hier den allgemeinen Fall darstellen, in dem man es mit Stichprobendaten zu tun hat.

vermindert sich der Steigungskoeffizient von $b = 0{,}80$ auf $b = 0{,}71$. Das heißt, der Industrialisierungsvariable wird nun ein etwas vermindertes Einflußgewicht zugeschrieben: Bei einer einprozentigen Zunahme der industriell Beschäftigten nimmt der SPD-Stimmenanteil im Schnitt nicht (mehr) um 0.8, sondern um 0.71 Prozentpunkte zu. Auch der Determinationskoeffizient vermindert sich von $r^2 = 0{,}50$ auf $r^2 = 0{,}37$. Das erscheint durchaus als sinnvoll, da zuvor der Industrialisierungsvariable ein Teil des Einflusses zugeschrieben wurde, die der Größe der Wahlkreise (bzw. weiterer mit ihr verbundener Faktoren wie Urbanisierung, Kommunikationsstruktur) zukommt[5].

10.1.1 Exkurs: Korrelation als Regression mit z-standardisierten Variablen (*)

Wir haben bereits in Teil I, Abschnitt 4.2.4 gezeigt, daß die klassische Definition des Korrelationskoeffizienten durch Pearson formal identisch ist mit der Kovarianz der z-standardisierten Variablen (siehe die dortige Gleichung (4-34)). Wir können nun zeigen, daß die Kovarianz z-standardisierter Werte identisch ist mit dem Steigungskoeffizienten der z-transformierten Variablen.

Die z-Transformation $z(x) = (x - \bar{x})/s_x$ und $z(y) = (y - \bar{y})/s_y$ führt dazu, daß die transformierte Variable ein arithmetisches Mittel von 0 und eine Standardabweichung bzw. Varianz von 1 aufweist (siehe Gleichung 7-9)). Wenden wir nun die Bestimmungsgleichung (10-8) für den Regressionskoeffizienten b auf die transformierte Variable an, so erhalten wir:

$$(10\text{-}17) \qquad b_{z(y)z(x)} = \frac{1/n\sum[z(x)_i - \overline{z(x)}][z(y)_i - \overline{z(y)}]}{1/n\sum[z(x)_i - \overline{z(x)}]^2}$$

$$= \frac{1}{n}\sum z(x)_i \cdot z(y)_i$$

Dieser Ausdruck läßt sich, wie in (4-34) gezeigt wurde und hier noch einmal wiederholt wird, zu der klassischen Definitionsformel von Pearson erweitern:

---

[5] In einem Vorgriff auf Kap. 11 läßt sich hier schon anmerken, daß in einer multiplen Regression, in der die Zahl der Wahlberechtigten neben dem Industrialisierungsgrad als weitere unabhängige Variable explizit in das Modell mit aufgenommen und auf eine Gewichtung verzichtet wird, für den Industrialisierungsgrad ein Steigungskoeffizient von $b = 0.72$ resultiert.

$$(10\text{-}18)\qquad \frac{1}{n}\sum z(x)_i \cdot z(y)_i = \frac{1}{n}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{1}{n}\cdot\frac{1}{s_x}\cdot\frac{1}{s_y}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - x)^2}\ \sqrt{\frac{1}{n}\sum(y_i - y)^2}}$$

$$= \frac{\text{Kovarianz } (y,x)}{s_x s_y} = r$$

Die z-Transformation ist eine lineare Transformation, die, Intervallskalenniveau vorausgesetzt, die Konfiguration der einzelnen Werte innerhalb der Punktwolke des Streudiagramms nicht verändert (siehe Abb. 10.6 im Vergleich zu Abb. 10.2). Da sich aber die Skaleneinheiten verändern, ändert sich die Neigung der Geraden. (Wegen des relativ groben Plot-Rasters können einzelne Werte auch ihre Position zueinander im Computerausdruck verschieben).

Die Interpretation der Korrelation als Regression z-standardisierter Werte macht deutlich: bei perfekter Korrelation, $r = |1|$, müssen nicht nur alle Punkte auf einer Geraden liegen (denn andernfalls wäre $r^2 < 1$), sondern diese Gerade muß identisch sein mit der Winkelhalbierenden in einem Koordinatenkreuz, auf dessen Achsen die z-standardisierten Werte abgetragen sind. Für die Punkte auf der Winkelhalbierenden gilt $z(x)_i = z(y)_i$, d.h. sie hat einen Steigungskoeffizienten von $b_{z(y)z(x)} = 1$.

Damit wird gezeigt, wie neben der PRE-Interpretation von r^2 auch Pearsons Korrelationskoeffizient r gedeutet werden kann: der Koeffizient r ist so konstruiert, daß er eine perfekte Korrelation nur für den Fall anzeigt, daß eine »relative« Abweichung der einzelnen x_i-Werte ($i = 1,2, \ldots, n$) von $\bar{x}$ (gemessen in Einheiten der Standardabweichung) bei jedem i mit einer gleich großen relativen Abweichung der y_i-Werte von $\bar{y}$ verbunden ist.

10.2 Theoretische Modellvoraussetzungen

Eine Regressionsanalyse kann nur durchgeführt werden, wenn

(a) eine Variable Y als »abhängige« Variable (»Kriteriumsvariable«) und eine oder mehrere andere Variablen als »unabhängige« Variablen X_1, X_2, ... , X_k (»Regressorvariablen«) definiert und gemessen sind und wenn

(b) die »funktionale Form« der Beziehung zwischen der Kriteriumsvariable Y und der Regressorvariable X (den Regressoren X_1, ..., X_k) festgelegt worden ist.

Das bedeutet, der Forscher muß auf Grund seines theoretischen Wissens (bzw. seiner Phantasie) ein »Modell« spezifizieren, das sich allgemein in Form einer Gleichung darstellen läßt:

(10-19)

$$Y = \alpha + \beta X + \epsilon \qquad \text{(bivariate R.)}$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \qquad \text{(multiple R.)}$$

Wir haben die Koeffizienten, die »Modellparameter«, jetzt mit griechischen Buchstaben bezeichnet, um damit anzuzeigen, daß wir uns nun auf der Ebene allgemeiner **theoretischer** Modelle bewegen, also Aussagen über (empirische oder hypothetische) **Populationen** machen wollen, statt lediglich einen als Stichprobe gegebenen Datensatz zu beschreiben. Die Regressionskoefizienten a und b, die wir mit Stichprobendaten ermitteln, dienen uns in der Regel nur als »Schätzer« für die theoretischen Modellparameter α und β.

Bei der »normalen« Regressionsanalyse formulieren wir ein Modell, das in seiner funktionalen Form »linear in den Parametern« ist (deshalb »lineare Regression«). Wir lassen Modelle zu, die »in den Variablen« nicht linear sind, aber durch Transformation linearisiert werden können (siehe ausführlicher hierzu Kap. 12.1), z. B.:

(10-20)

$$Y = \alpha + \beta X^2 \Rightarrow y = \alpha + \beta X^* \,,\; X^* = X^2$$

$$Y = \alpha X^{\beta} \cdot 10^{\epsilon} \Rightarrow \log Y = \log\alpha + \beta \log X + \epsilon$$

$$Y^* = \alpha^* + \beta X^* + \epsilon$$

Wir schließen Modelle aus wie

(10-21) $Y = \alpha X^{\beta} + \epsilon$ (man beachte die additive, nicht multiplikative Fehlerkomponente)

Solche Modelle sind auch durch eine Transformation der Variablen (z. B. Logarithmieren) nicht in eine lineare Form zu bringen.

Normalerweise möchte der Forscher mehr erreichen, als einen gegebenen Datensatz zu beschreiben und die Werte einer Variablen Y mit Hilfe der Werte einer anderen Variablen X optimal zu prognostizieren. Er möchte die korrelativen Zusammenhänge, die er beobachtet, nicht als bloße zeitliche Koinzidenz interpretieren, sondern als Ausdruck einer »strukturellen«, letztlich »kausalen« Beziehung. Was mit einer **Kausal**hypothese begrifflich gemeint ist, läßt sich am ehesten in Form eines irrealen Konditionalsatzes ausdrücken: Wenn man über die festgestellte Koinzidenz hinaus einen kausalen Zusammenhang behauptet, nimmt man an, daß Y auch dann mit X kovariieren würde, wenn die x_i-Werte sich nicht unabhängig vom Beobachter eingestellt hätten, sondern bewußt (experimentell) erzeugt worden wären (zu dieser Kausalitätskonzeption siehe von Wright 1974, S. 32, 72 ff.).

Wenn man die Steigungskoeffizienten im Kontext einer solchen allgemeinen (Kausal-)Theorie interpretieren will, werden sie zu Indikatoren für den **spezifischen** Einfluß des jeweiligen Regressors auf die abhängige Variable. Der spezifische Einfluß eines Regressors läßt sich aber nur ermitteln, wenn alle »relevanten« Einflußfaktoren im Modell (in der Regressionsgleichung) erfaßt worden sind - was bei einer bivariaten Analyse ziemlich unwahrscheinlich ist. Deshalb benötigen wir in der Regel die multiple Regression (siehe Kap.11) oder noch komplexere Modelle (sog. »Pfadmodelle« oder »Strukturgleichungsmodelle«).

Bei der **Prognose** der y_i-Werte stört es weniger, wenn einem bestimmten Regressor X_k ($k = 1,2, \ldots, K$) durch die Methode der kleinsten Quadrate vielleicht nur deshalb ein hohes Gewicht zugewiesen wird, weil X_k wie auch Y mit einer anderen Einflußvariable X^* korreliert, die (leider) nicht im Modell berücksichtigt wurde. Aber als Indikator für die spezifische kausale Einflußstärke von X_k ist der Steigungskoeffizient b_k in diesem Falle irreführend. Wie wir unten in Ziff. 2 noch erläutern werden, ist es bei diesem Problem von entscheidender Bedeutung, ob die ausgelassene Variable nur mit Y oder auch mit dem berücksichtigten Regressor korreliert.

Der Fehlerterm ε repräsentiert also nicht nur »Fehler«, die bei der Datenerhebung gemacht werden (Stichproben- und Meßfehler), sondern auch diejenigen Einflußfaktoren, die Y (mit)bestimmen, aber im Regressionsmodell nicht berücksichtigt wurden - sei es, weil sie nicht bekannt sind, sei es, weil sie aus technischen oder finanziellen Gründen nicht erhoben worden sind. Man bezeichnet sie auch als »implizite Variablen«.

Durch die Berücksichtigung des »Fehlers« bzw. der »impliziten Variablen«, also mit dem Wechsel von Gleichung (10-1) zu (10-3) bzw. (10-19) vollziehen wir den Übergang von einem **deterministischen** zu einem **stochastischen** Modell, einem Modell also, in dem eine Zufallskomponente ε ausdrücklich berücksichtigt wird. Wir beanspruchen lediglich, mit unserem Modell die bedingten Erwartungswerte $E(Y \mid x_i)$, nicht die individuellen y_i-Werte prognostizieren zu können. Die Fehlergrößen streuen »zufällig«, das heißt aus unbekannten oder quantitativ nicht erfaßten Gründen um die Regressionsgerade. (Wie wir noch sehen werden, müssen dennoch bestimmte Annahmen über das »Verhalten« dieser Fehler gemacht werden.) Der Einfluß der Regressorvariablen wird im Unterschied dazu als »systematisch« bezeichnet[6].

Es gibt noch weitere Modellvoraussetzungen, die erfüllt sein müssen, wenn man die Regressionskoeffizienten anhand von Stichprobendaten zuverlässig schätzen, testen und im Sinne einer Kausalhypothese interpretieren will. Wir wollen sie hier einmal insgesamt (einschließlich der bereits genannten) auflisten, beschränken uns dabei aber auf eine sehr gedrängte Darstellung. Zu diesem Zeitpunkt wird weniger ein vollständiges Verständnis der einzelnen Voraussetzungen, eher ein genereller Einblick in die Bedingtheit der regressionsanalytischen Ergebnisse angestrebt. Damit soll allzu naiven oder leichtfertigen Interpretationen vorgebeugt werden, zu denen ansonsten die leichte Verfügbarkeit der EDV verleiten mag. (Der Leser mag die folgenden Punkte zunächst nur flüchtig durchsehen und nach der Lektüre von Abschn. 10.3 zu ihnen zurückkehren.)

Die Erfüllung dieser Modellvoraussetzungen soll sicherstellen, daß die Regressionsparameter konsistent, effizient und erwartungstreu geschätzt werden (siehe Kap. 8.3). Um Konfidenzintervalle schätzen und Signifikanztests durchführen zu können, müssen die Varianzen bzw. Standardfehler der entsprechenden Koeffizienten bekannt sein oder ebenfalls aus den vorliegenden Daten geschätzt werden. Für die Schätzung der Standardfehler gelten natürlich die gleichen Gütekriterien wie für die Schätzung der Regressionskoeffizienten, aber sie werden nicht unbedingt unter den gleichen Voraussetzungen eingelöst. Es kann z. B. vorkommen, daß unter bestimmten Bedingungen zwar die Regressionsparameter selbst (»Punktschätzer«) erwartungstreu geschätzt werden, nicht aber ihre Standardfehler.

Im einzelnen sind folgende Punkte zu beachten:

---

[6] Mit dem Begriff »Zufall« ist hier lediglich unser Nichtwissen eingestanden, keine philosophische These über die letzthinnige Determiniertheit oder Zufälligkeit des Weltgeschehens verkündet. Ob man »stochastische« oder »deterministische« Modelle konstruiert, ist also keine Frage der Weltanschauung, sondern der Forschungspragmatik.

1. Die funktionale Form der Beziehung zwischen der Y-Variable und dem oder den Regressoren muß korrekt spezifiziert sein. Die Variablen müssen u. U. so transformiert werden, daß die Linearitätsvoraussetzung erfüllt ist (siehe Kap. 12.1) - oder man muß eine Variante der nicht-linearen Regression (siehe Kap. 12.2) wählen.
2. Die Werte $X = x_i$ eines Regressors müssen
   a) entweder vom Forscher selbst experimentell festgelegt oder aus anderen Gründen perfekt vorhersagbar sein (der Regressor X wäre dann keine Zufallsvariable) oder, falls dies nicht zutrifft, d. h.
   b) falls X eine Zufallsvariable darstellt, müssen die x_i unabhängig von den Fehlerwerten ε_i sein (siehe unten).

Nur wenn eine der beiden Voraussetzungen erfüllt ist, können die mit Stichprobendaten ermittelten Regressionskoeffizienten optimale (effiziente, erwartungstreue und konsistente) Schätzer für die Regressionskoeffizienten des theoretischen Modells (für die Population) sein. Wenn statt der Voraussetzung 2 b) lediglich die Bedingung efüllt ist, daß die x_i und ε_i nicht **linear** miteinander korrelieren, bleiben die optimalen Schätzeigenschaften nur »asymptotisch« erhalten, werden also nur näherungsweise bei großen Stichproben (Faustregel $n > 100$) erreicht.

Mit einer Korrelation zwischen den x_i und den ε_i ist vor allem dann zu rechnen, wenn ein relevanter Regressor in der geschätzten Regressionsgleichung nicht berücksichtigt wurde. »Relevant« in diesem Sinne ist eine ausgelassene Variable, wenn sie sowohl mit der abhängigen Variable als auch mit der oder den in der Gleichung berücksichtigten Regressoren korreliert. Dieser Fall eines sog. Spezifikationsfehlers führt dazu, daß die geschätzten Regressionskoeffizienten keine erwartungstreuen Schätzer der »wahren« Regressionsparameter sind; sie können sowohl unterals auch überschätzt sein, je nachdem, wie der oder wie die ausgelassenen Regressoren mit dem oder den berücksichtigten Regressoren (positiv oder negativ) korrelieren. Wenn relevante Regressoren ausgelassen worden sind, wird den in der Gleichung berücksichtigten Regressoren bei der Minimierung der Fehlerquadratsumme ein Teil des Einflußgewichts (des Steigungskoeffizienten) zugewiesen, das den nicht berücksichtigten Variablen zukommt[7]. Falls die ausgelassene Variable X^* nur mit Y nicht aber mit einem berücksichtigten Regressor X_k korreliert, wird dessen Steigungskoef-

[7] Bei falsch spezifizierter Form der Beziehung und beim Auslassen relevanter Variablen entstehen Schätzprobleme auch dann, wenn man es nicht mit Stichprobendaten, sondern mit Populationsdaten zu tun hat; denn die errechneten Regressionskoeffizienten sind stets abhängig von dem vorgegebenen Modell.

fizient b_k unverzerrt (erwartungstreu) geschätzt. Allerdings entsteht auch in diesem Falle ein Problem dadurch, daß der Standardfehler für b_k überschätzt wird, so daß beim Signifikanztest leichter ein Fehler zweiter Art auftreten kann. Das heißt, der Steigungskoeffizient b_k wird ohne Berücksichtigung von X^* eher als »insignifikant« zurückgewiesen, als wenn X^* in der Regressionsgleichung mit berücksichtigt worden wäre.

Es ist ziemlich schwierig, in den Daten selbst Hinweise zu finden, ob man relevante Variablen ausgelassen hat oder nicht. Der Gedanke, doch einfach die **beobachteten** Fehler e_i mit den x_i zu korrelieren, führt nicht weiter, da durch das Kleinstquadratverfahren $r_{ex} = 0$ erzwungen wird, auch wenn die »wahren« Fehler ε_i mit den x_i korrelieren (siehe Hanushek/Jackson 1977, S. 51). Einige multivariate Analyse - bzw. Schätzverfahren, die wir in diesem Skript nicht behandeln, bieten gewisse Möglichkeiten zu testen, ob relevante Variablen ausgelassen wurden. Gewisse Hinweise kann man u.U. den Residuenplots entnehmen (siehe Abschn. 10.4). Die korrekte Modellspezifikation ist jedoch vor allem eine Sache des theoretischen Wissens. Diese Problemlage sollte nicht zu dem Schluß verleiten, es sei ratsam, möglichst viele Regressoren in die Gleichung einzubauen, selbst wenn die substanzwissenschaftliche Theorie keine guten Gründe dafür liefert. Auch irrelevante Variablen, die keinerlei Einfluß auf Y haben, führen zu unliebsamen Konsequenzen bei der Schätzung der Modellparameter. Zwar werden die Regressionskoeffizienten dadurch nicht erwartungsuntreu, aber ihre Standardfehler werden vergrößert, wenn der Regressor mit einer irrelevanten Variablen korreliert; sie sind also nicht mehr effizient und können im Einzelfall von ihrem Zielwert weit abweichen. Immerhin wird dieser (vergrößerte) Standardfehler unverzerrt geschätzt, so daß die Signifikanztests valide bleiben.

3. Die Variablen sollen fehlerfrei gemessen sein. Diese Voraussetzung ist in der sozialwissenschaftlichen Forschung praktisch nicht erfüllbar. Das liegt nicht nur an technischen Unzulänglichkeiten des Messens, sondern auch daran, daß die Variablen häufig theoretische Konstrukte darstellen, die über empirische Indikatoren nur indirekt gemessen werden.

   Das Meßfehlerproblem ähnelt formal dem der Spezifikationsfehler, da es auch hier um unkontrollierte Einflußfaktoren geht, die im Modell nicht berücksichtigt sind. Im allgemeinen unterscheidet man zufällige (»random«) und systematische, nicht-zufällige (»non-random«) Meßfehler. Als zufällig bezeichnet man Meßfehler dann, wenn sie
   a) sich in der Summe bei jeder Variable ausgleichen,

b) wenn die einzelnen Fehler untereinander und vom wahren Wert unabhängig sind,
c) wenn die Meßfehler der einen Variablen unabhängig von den Meßfehlern der anderen Variablen sind.

Bei den »systematischen« Meßfehlern ist mindestens einer dieser Punkte nicht erfüllt. Solange man sie nicht kennt und nicht als »Hilfstheorie« in das Regressionsmodell integriert, beeinflussen sie das Regressionsergebnis in nicht angebbarer Weise. Die Folgen zufälliger Meßfehler lassen sich hingegen benennen und unter bestimmten Voraussetzungen bei der Schätzung der Regressionskoeffizienten (und des Determinationskoeffizienten) berücksichtigen (in komplexeren Verfahren, die wir hier nicht erklären).
Zufällige Meßfehler in der abhängigen Variablen berühren nicht die Erwartungstreue der Regressionskoeffizienten. Höhere Meßfehler bedeuten allerdings eine höhere Residualvarianz. Dadurch wird der Determinationskoeffizient (folglich auch der Korrelationskoeffizient) gemindert. Außerdem wird, wie wir im nächsten Abschnitt noch sehen werden, der Standardfehler der Regressionskoeffizienten erhöht, d. h., der Schätzer wird weniger effizient.
Zufällige Meßfehler in der unabhängigen Variablen hingegen lassen den Regressionsschätzer selbst inkonsistent werden. Im bivariaten Fall wird der Koeffizient unterschätzt (negativer »bias«). Im multivariaten Falle kann man das nicht so allgemein sagen, weil die Korrelationen zwischen den einzelnen Regressoren hierfür bedeutsam sind. Es kann sowohl zu Unter- als auch zu Überschätzungen kommen.
Die Größe eines Meßfehlers läßt sich schätzen, wenn eine Variable mit mehreren Indikatoren gemessen wird. In den Sozialwissenschaften fordert man zunehmend, demgemäß zu verfahren und bei der empirischen Überprüfung substantieller Theorien Meßmodelle in das zu überprüfende Gesamtmodell zu integrieren. Ein methodischer Ansatz hierzu ist unter dem Namen LISREL prominent geworden (siehe z. B. Pfeifer/Schmidt 1987).

4. Die Fehler insgesamt (nicht nur die Meßfehler), so wird weiter vorausgesetzt, streuen mit konstanter Varianz um den Erwartungswert Null: $E(\varepsilon_i^2)=\sigma^2$, $E(\varepsilon_i) = 0$ für alle i. Die Annahme, daß die Erwartungswerte der Fehler konstant sind, impliziert u. a. die Voraussetzung, daß die funktionale Form der Regressionsgleichung korrekt spezifiziert wurde (siehe oben, Ziff. 1). Daß diese konstanten Erwartungswerte gleich Null sind, stellt sicher, daß nicht nur der Steigungskoeffizient, sondern auch der Ordinatenabschnitt erwartungstreu geschätzt werden kann (siehe Hanushek/Jackson 1977, S. 51, 71, 136).

Die Voraussetzung konstanter (»homogener«) Fehlervarianzen (»Homoskedastizität«) ist in unserem Analysebeispiel (siehe das Streudiagramm in Abb. 10.2) vermutlich nicht erfüllt (»Heteroskedastizität«). Es sieht so aus, als ob die Varianz mit wachsendem Industrialisierungsgrad zunehme (größere Varianzen der Y-Werte bedeuten auch größere Varianzen der Fehler).

Inhomogene Varianzen können aus unterschiedlichen Gründen vorliegen. Sie können z. B. durch Meßfehler oder durch nicht berücksichtigte Regressoren verursacht sein, die mit den berücksichtigten Regressoren korrelieren (siehe oben, Ziff. 2). Bei Aggregatdaten (wie in unserem Beispiel) ist stets mit Streuungsungleichheit zu rechnen. Das ergibt sich schon allein dadurch, daß die Aggregate oft eine ungleiche Zahl von Fällen (hier Einwohner in Wahlbezirken) zusammenfassen und Varianzen von Stichprobenfunktionen von den Fallzahlen (bzw. Freiheitsgraden) abhängen. Die Varianzen von Prozentanteilen sind außerdem noch abhängig von dem jeweiligen Erwartungswert (siehe Gleichung 7-4); sie sind um so geringer, je weiter die Erwartungswerte (Prognosewerte) von der 50-Prozent-Marke entfernt sind.

Die verschiedenen Faktoren, die Heteroskedastizität hervorrufen, können sich in ihrer Wirkung addieren oder überlagern.

Die Schätzung der Regressionskoeffizienten bei Heteroskedastizität nach der üblichen Kleinstquadratmethode bleibt zwar erwartungstreu und konsistent, aber die Standardfehler erhöhen sich: die Schätzer sind nicht effizient, nicht einmal asymptotisch (bei größer werdender Stichprobe) effizient. Darüber hinaus wird diese vergrößerte Varianz der geschätzten Regressionskoeffizienten ihrerseits unter der Heteroskedastizitätsbedingung nicht erwartungstreu geschätzt. Wenn die Varianz mit den X-Werten positiv korreliert (wie möglicherweise in unserem Beispiel), wird der Standardfehler unterschätzt. Unterschätzte Standardfehler führen dazu, daß die Konfidenzintervalle enger werden, als sie es bei der gewählten Irrtumswahrscheinlichkeit α sein dürften. Somit wird das tatsächliche Fehlerrisiko für die Ablehnung der Nullhypothese größer als durch den Alpha-Wert angezeigt.

Form und Ausmaß der Varianzheterogenität versucht man bei der Residuenanalyse (siehe Abschn. 10.4) zu ermitteln. (Es gibt auch formale Testmöglichkeiten: siehe Kmenta 1971, S. 267 f.).

Dem Praktiker zum Trost zitieren Berry/Feldman (1985, S. 78) eine Arbeit von Bohrnstedt und Carter, die zu dem Schluß gelangen, »that unless heteroscedasticity is 'marked', significance tests are 'virtually unaffected', and thus OLS can be used without concern of serious distortion«; aber, fahren sie fort, »in some analyses, heteroscedasticity may be severe«.

5. Die Fehler ε_i dürfen nicht nur mit dem oder den Regressoren nicht korrelieren (siehe oben, Ziff. 2), sie dürfen auch nicht untereinander korrelieren, nicht »autokorrelieren«. Autokorrelationen können (z. B. auf Grund von Meßfehlerprozessen oder räumlicher Nähe von Erhebungseinheiten) auch in Querschnittdaten auftreten. Sie sind dort aber normalerweise nicht identifizierbar, weil die Fälle in der Datenanalyse beliebig angeordnet werden können. Bei Zeitreihendaten sind autokorrelierte Fehler regelmäßig zu erwarten. Dort können sie auch empirisch identifiziert werden, weil die Meßergebnisse durch ihren Zeitindex eindeutig geordnet sind.
   Die Konsequenzen autokorrelierter Fehler entsprechen in etwa denen heterogener Varianzen: Die Regressionsschätzer bleiben erwartungstreu, werden aber ineffizient. Ihr Standardfehler wird erwartungsuntreu und inkonsistent geschätzt. Erwartungsuntreu wird auch die Schätzfunktion für die Residualvarianz.
   Für die Bearbeitung dieser Probleme sind verschiedene Lösungsstrategien vorgeschlagen worden, z. B. die Verallgemeinerte Kleinstquadratmethode (GLS - »Generalized Least Squares«) oder die Box/Jenkins-Methode der Zeitreihenanalyse.

6. Die Fehler sollen normalverteilt sein. Diese Annahme kann man mit entsprechenden Diagrammen leicht überprüfen (siehe Abschn. 10.4). Die Normalverteilungsannahme ist in den meisten Anwendungsfällen praktisch nicht relevant. Ist sie nicht erfüllt, hat das keine Auswirkungen auf die (Punkt-) Schätzung der Regressionskoeffizienten. Man benötigt aber die Annahme normalverteilter y_i, um die Normalverteilung der Regressionskoeffizienten theoretisch begründen zu können und auf dieser Basis Konfidenzintervalle schätzen und Signifikanztests durchführen zu können. Allerdings läßt sich auch hier der zentrale Grenzwertsatz anwenden: Die Verteilung der Regressionskoeffizienten nähert sich bei $n \to \infty$ dem Modell der Normalverteilung an, auch wenn die y_i nicht normalverteilt sind (Kmenta 1971, S. 248). Bei kleinen Stichproben sind die Intervallschätzungen und Signifikanztests aber nur valide, wenn die y_i normalverteilt sind. Bei den F-Tests im Rahmen der multiplen Regression (siehe Kap. 11) ist ebenfalls darauf zu achten, daß die Normalverteilungsannahme in etwa erfüllt ist. Zu beachten ist außerdem, daß die Effizienz der Kleinstquadrate-Schätzer stark gemindert sein kann, wenn in der Fehlerverteilung erheblich mehr extreme Werte auftauchen als nach der Normalverteilungsannahme zu erwarten wären. Für diesen Fall werden in der Literatur sog. **robuste** Schätzverfahren empfohlen (siehe Schlittgen 1987, S. 389, 393 f.)

7. Falls die Regressionskoeffizienten im Sinne eines **kausalen** Zusammenhangs zwischen den Variablen interpretiert werden sollen, müs-

sen sich die Daten zum Zeitpunkt der Messung im Gleichgewicht (»Äquilibrium«) befinden; sie dürfen sich nicht, z. B. auf Grund eines externen Ereignisses, zum Zeitpunkt der Messung in rascher Bewegung zu einem neuen Gleichgewichtszustand befinden. Der Schätzung des **Steigungskoeffizienten** ist ein eventuell vorliegendes Disäquilibrium aber nur dann abträglich, wenn die Veränderungsrate der y_i mit den x_i korreliert. In unserem Beispiel ist nicht auszuschließen, daß sowohl die SPD-Stimmenanteile als auch das Industrialisierungsniveau in den einzelnen Wahlbezirken in Bewegung sind. Vermutlich sind die Veränderungsraten in den einzelnen Wahlbezirken sehr unterschiedlich. Man wird vielleicht auch annehmen wollen, daß die Veränderungsrate der SPD-Präferenz mit dem erreichten Industrialisierungsniveau kovariiert. Das würde bedeuten, daß Daten aus einem anderen Erhebungszeitpunkt zu einem anderen Steigungskoeffizienten führen müßten. Die Steigungskoeffizienten wären also zeitabhängig; die strukturelle Beziehung würde durch einen einzelnen Koeffizienten nicht adäquat angezeigt. Leider ist die Industrialisierungsvariable nicht mehrmals erhoben worden, so daß wir die Stabilität der Regressionskoeffizienten nicht überprüfen können.

8. Wenn wir ein Regressionsmodell mit mehreren Regressorvariablen $X_1, \ldots, X_K$ spezifiziert haben, dürfen zwischen ihnen keine perfekten linearen Zusammenhänge bestehen. Schon bei hoher »Multikollinearität« entstehen Probleme, weil sich die Standardfehler erhöhen, die Schätzwerte also dazu tendieren, relativ weit vom wahren Wert abzuweichen (siehe Kap. 11.2.2).

10.3 Intervallschätzung und Signifikanztest

Wenn wir das Ergebnis unserer Regressionsanalyse nicht über die vorliegenden Untersuchungseinheiten hinaus verallgemeinern, die Regressionsgleichung nicht als theoretisches Modell interpretieren wollen, entfällt die Notwendigkeit der Berechnung von Konfidenzintervallen und Signifikanztests. Zumindest der Übung wegen wollen wir unsere Daten jedoch als Stichprobendaten behandeln und deshalb die entsprechenden Analyseschritte erläutern.

In ähnlicher Weise wie der Standardfehler des arithmetischen Mittels theoretisch abgeleitet werden kann (siehe Abschn. 7.3), läßt sich auch der Standardfehler des Regressionskoeffizienten theoretisch bestimmen: Wenn die im vorigen Abschnitt genannten Voraussetzungen erfüllt sind, ist der Steigungskoeffizient b als Schätzer des Modellparameters ß normalverteilt um den Erwartungswert ß mit dem Standardfehler

$$(10\text{-}22) \qquad \sigma_b = \frac{\sigma_\varepsilon}{\sqrt{\sum(x_i-\bar{x})^2}}$$

Wir schreiben also

$$b \sim N\left(\beta, \frac{\sigma}{\sqrt{\sum(x_i-\bar{x})^2}}\right)$$

Da die wahre Fehlervarianz σ^2_ε nicht bekannt ist, muß die Standardabweichung aus den Stichprobendaten mit

$$(10\text{-}23) \qquad \hat{\sigma}_\varepsilon = s_e = \sqrt{\frac{1}{n-2}\Sigma(y_i-\hat{y}_i)^2}$$

geschätzt werden. Die Zahl der Freiheitsgrade, n-2, ergibt sich aus der Zahl der Fälle minus der Zahl der im Regressionsmodell geschätzten Parameter. Die Schätzung bedeutet, wie schon beim arithmetischen Mittel (siehe Abschn. 7.4.2), daß b nun nicht mehr normal-, sondern t-verteilt ist. Allerdings gilt auch hier, daß sich die t-Verteilung schon bei 30 Freiheitsgraden (Faustregel) der Normalverteilung weitgehend angenähert hat. Die folgenden Formeln schreiben wir aber so, als würden wir in jedem Falle die t-Verteilung anwenden.

In unserem Beispiel wird für den Steigungskoeffizienten b=0,71 ein Standardfehler von $s_b = 0{,}047$ ausgewiesen. Da wir unsere Daten wie Stichprobendaten behandeln, wollen wir nun das Konfidenzintervall ausrechnen und Signifikanztests durchführen. Wenn man die bereits zitierten SPSS^x-Kommandos zur Regressionsanalyse um den Befehl /STATISTICS=CI (einzufügen vor DEPENDENT) ergänzt, werden auch das 95 %-Konfidenzintervall sowie der kritische t-Wert für einen zweiseitigen Signifikanztest bei einem Fehlerrisiko von $\alpha \leqq 0{,}05$ mit ausgedruckt (siehe Abb. 10.7). Der Übung wegen wollen wir deren Berechnung hier noch

einmal Schritt für Schritt durchgehen und dabei den Ergebnisausdruck erläutern.

Der Standardfehler des Steigungskoeffizienten resultiert laut Gleichung (10-22) aus zwei Komponenten, der Standardabweichung der Residuen im Zähler und der Variation des Regressors im Nenner. Da die Variation mit der Zahl der Fälle wächst, folgt daraus, daß der Standardfehler von b um so kleiner wird, je größer der Stichprobenumfang ist. Die Standardabweichung der Residuen (Zählerkomponente) muß gemäß Gleichung (10-23) geschätzt werden. In der Ergebnisliste des Computerausdrucks ist dieser Schätzer als STANDARD ERROR angegeben. Wir können seine Berechnung mit Hilfe weiterer Angaben aus Abb. 10.7 (unter »Analysis of Variance«) nachvollziehen:

$$(10\text{-}24) \qquad s_e = \sqrt{\frac{77406{,}27}{393}} = \sqrt{196{,}96} = 14.034$$

Die Varianz der Regressorvariablen (Industrialisierungsindikator) kann über das DESCRIPTIVE-Subkommando ermittelt werden. In unserem Beispiel ist

$$(10\text{-}25) \qquad \hat{\sigma}_x^2 = 1/(n-1)\sum(x_i-\bar{x})^2 = 230{,}16$$

Somit erhalten wir für den Nennerausdruck in (10-22)

$$(10\text{-}26) \qquad \sqrt{\sum(x_i-\bar{x})^2} = \sqrt{(230{,}16 \cdot 394)} = 301{,}14$$

Daraus ergibt sich der bereits erwähnte Schätzer für den Standardfehler des Steigungskoeffizienten

$$(10\text{-}27) \qquad \hat{\sigma}_b = \frac{s_e}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{14{,}03}{301{,}14} = 0{,}047$$

Dieser Wert kann ebenfalls als »SE B« direkt dem Ergebnis-Ausdruck in Abb. 10.7 entnommen werden. Bei einer Irrtumswahrscheinlichkeit von $\alpha \leqq 0.05$ erhalten wir für das Konfidenzintervall nach Gleichung (8-23) die kritischen t-Werte (siehe Anhang A, Tab. 2)

(10-28) $t_{n-2;\alpha/2} = -1,97$

$t_{n-2;1-\alpha} = +1,97$

Daraus ergibt sich das 95-Prozent-Vertrauensintervall

(10-29)

$$P(b - t_{\alpha/2} \cdot \hat{\sigma}_b \leq \beta \leq b + t_{\alpha/2} \cdot \hat{\sigma}_b) = 0,95$$

$$P(0,713 - 1,97 \cdot 0,047 \leq \beta \leq 0,713 + 1,97 \cdot 0,047) = 0,95$$

$$P(0,62 \leq \beta \leq 0,81) = 0,95$$

Diese Intervallgrenzen stimmen bis auf Rundungsfehler mit den Werten überein, die im Ergebnisausdruck unter der Bezeichnung »95 % CONFDNCE INTRVL B« erscheinen. (Eine Zeile tiefer findet man auch die entsprechenden Angaben für den Ordinatenabschnitt, »CONSTANT«)

Da der Wert Null außerhalb dieses Intervalls liegt, ist auch schon klar, daß die Nullhypothese H_o: $\beta = 0$ mit einem Fehlerrisiko von $\alpha \leqq 0,05$ zurückgewiesen werden kann.

Der **einseitige** Signifikanztest mit

H_1: $\beta > 0$
H_o: $\beta \leqq 0$, $\alpha \leqq 0,05$

wird dann erst recht zur Ablehnung der Null-Hypothese führen. Der Übung wegen sei er hier noch einmal durchgespielt: Der kritische t-Wert ist nun

(10-30) $t_{n-2;1-\alpha} = t_{n-2;1-0,05} = 1,65$

Der beobachtete Regressionskoeffizient $b = 0,71 = \beta$ liegt aber mehr als nur 1,65 Standardabweichungen von seinem Erwartungswert $E(b) = 0$ entfernt:

(10-31) $$t^* = \frac{b-0}{\hat{\sigma}_b} = \frac{0,71}{0,047} = 15,1$$

Das empirische Signifikanzniveau α^* hierfür ist kleiner als ein Zehntausendstel. (Im Computerausdruck ist das empirische Signifikanzniveau für

den zweiseitigen Test bis auf vier Stellen hinter dem Komma angegeben.) Der Regressionskoeffizient ist also, statistisch gesehen, hoch signifikant.

Signifikantstests und die Berechnung von Konfidenzintervallen beruhen auf Modellannahmen über das Verhalten der Fehler (siehe Abschn. 10.2), denen die Praxis nur selten voll entspricht. Deshalb sind alternative Berechnungsmethoden vorgeschlagen worden, z. B. das sog. »Jackknifing«. Eine knappe Einführung hierzu gibt Achen (1982, Kap. 4). Dort findet der Leser auch Überlegungen, die den konventionellen Signifikanztest problematisieren und andere Entscheidungskriterien (»substantielle« Signifikanz) ins Spiel bringen.

Zum Ergebnisausdruck in Abb. 10.7 ist noch anzumerken, daß er terminologisch auf die multiple Regression, also die Regression mit mehreren unabhängigen Variablen bezogen ist. Deshalb ist auch bei nur einer Regressorvariablen die Rede von »Multiple R«. Das »Adjusted R Square«, den mit »Beta« überschriebenen Koeffizienten sowie den Varianzanalytischen F-Test, werden wir in Kap. 11 erläutern.

10.4 Überprüfung von Modellvoraussetzungen: Residuenanalyse und gewichtete Regression

Die Analyse der Residuen ist das wichtigste Hilfsmittel, mit dem man wenigstens einige der in Abschn. 10.2 genannten Modellvoraussetzungen empirisch prüfen kann. Jede Art von Regelmäßigkeit in den Residuen deutet auf die Verletzung irgendeiner Voraussetzung hin. In der Forschungspraxis stützt man sich dabei vor allem auf die visuelle Inspektion verschiedener Diagramme:

(a) des Streudiagramms zwischen den Residuen e_i (meistens in standardisierter Form) und den Prognosewerten $\hat{y}_i$ sowie den Regressorwerten x_i,
(b) eines Histogramms der Residuen und eines sog. »Normal Probability Plots«, in dem die kumulierte Häufigkeitsverteilung der Residuen und der Standardnormalverteilung gegeneinander geplottet werden,
(c) eines Plots der Residuen gegen die Sequenz der Fälle.

Wir wollen hier nur einige Standardverfahren skizzieren, ohne die Feinheiten der Residuenanalyse zu erläutern (ausführlichere bzw. weiterführende Darstellungen geben Draper/Smith 1981, Kap. 3; Cook/Weisberg 1982; Norusis 1985, Kap. 2.16 - 2.31, 2.49; Belsley et al. 1980). In SPSS^x stehen eine Reihe von Optionen zur Residuenanalyse zur Verfügung, wobei wir hier folgende Auswahl treffen (diese Subkommandos werden nach /DEPENDENT eingegeben):

```
/RESIDUALS = HISTOGRAM(SRESID) NORMPROB
             OUTLIERS ID(WKNR)
/CASEWISE = DEFAULT ALL
/SCATTERPLOT = (*RESID,*PRED)
             = (*SRESID,*PRED) (*SRESID,INDUSTRY)
/SAVE PRED(PREDSPD) RESID(RES) SRESID(STUDRES)
PLOT PLOT RES STUDRES WITH INDUSTRY PREDSPD
```

Die Residuen können auf zweierlei Weise standardisiert sein. Vertraut ist uns bereits die z-Standardisierung (siehe Teil I, Abschn. 4.2.4 oder den Exkurs 10.1.1), die die Residuen auf eine Standardabweichung von 1 und ein arithmetisches Mittel von 0 »normiert«. Eine weitere Standardisierung führt zu den sog.»studentized residuals« (SRESID). Hier wird der Residualwert durch einen Schätzer seines Standardfehlers dividiert, der selbst wiederum mit den X-Werten variiert. Dahinter steht ein formales Element des Regressionsmodells, das wir hier nicht näher erläutern wollen: In das Modell ist eine Tendenz eingebaut, die Varianzen der geschätzten (nicht der »wahren«[8]) Residuen um so kleiner werden zu lassen, je weiter die jeweiligen X-Werte von ihrem arithmetischen Mittel entfernt liegen (siehe Schlittgen 1987, S.393). Die s-Standardisierung schaltet diesen Einfluß aus und erlaubt somit eine angemessenere Prüfung der Frage, ob eine Heteroskedastizität vorliegt, die durch Modelldefekte oder/und Meßfehler verursacht sein könnte. In der Praxis ergeben sich aber selten nennenswerte Unterschiede in den Plots der z- oder s-standardisierten Residuen und auch kaum welche in den Plots standardisierter und nicht-standardisierter Residuen.

Die entsprechenden SPSS^x-Subkommandos zielen in der Voreinstellung auf die z-standardisierten Werte. In dem obigen Subkommando /RESIDUALS fordern wir für das Histogramm die »studentized« und für den »Normal Probability Plot« (siehe unten), die z-standardisierten Residuen an (zwischen ihnen besteht in unserem Beispiel kaum ein Unterschied). Mit den OUTLIERS werden uns die zehn extremsten Werte (hier wieder in z-standardisierter Form) mit Angabe der Wahlkreisnummer (WKNR) ausgedruckt. Das /CASEWISE-Kommando mit seinen weiteren Spezifikationen liefert uns einen Plot aller Residuen pro Wahlkreisnummer nebst weiteren Angaben wie z. B. der numerischen Größe der Residuen. Mit dem /SCATTERPLOT-Befehl erhalten wir verschiedene Streudiagramme, wobei die jeweils erstgenannte Variable in den Klammerausdrücken die Ordinaten-Variable darstellt. Variablen, die wie PRED (die vom Modell generierten Prognosewerte für die abhängige Variable) systemintern er-

[8] Die »wahren« Residuen sind die Abweichungen von der »wahren« Regressionslinie, die man bei korrekter Modellspezifikation ermitteln könnte, wenn die Populationsdaten vollständig vorlägen.

zeugt werden, müssen mit einem Sternchen-Symbol (*) markiert sein. Vor dem Plotten werden alle Variablen automatisch z-standardisiert, auch wenn nicht *ZRESID, sondern *RESID verlangt wird. Will man Streudiagramme mit unstandardisierten Werten erstellen, müssen sie zunächst mit dem /SAVE-Kommando gesichert werden. Dabei muß den systemerzeugten Variablen (jetzt ohne Sternchen) in Klammern ein gleicher oder ein davon abweichender Name zugewiesen werden, der in späteren Kommandos, z. B. dem PLOT-Kommando benutzt wird.

Wir beginnen nun die Residuenanalyse mit verschiedenen Residuenplots in Abb. 10.8: a) der nicht standardisierten Residuen gegen den Indikator INDUSTRY der industriellen Entwicklung, b) der s-standardisierten Residuen gegen INDUSTRY c) der s-standardisierten Residuen gegen die z-standardisierten Prognosewerte (*PRED), d) der z-standardisierten Residuen gegen *PRED.

Die verschiedenen Streudiagramme weichen in ihrem Muster nur geringfügig voneinander ab. Sie bestätigen den Eindruck von Abb. 10.2. Zum ersten nimmt die Menge der stark negativen Residuen mit höheren X-Werten zu. Wenn wir davon absehen, dominiert in allen Diagrammen das Bild einer trichterförmigen Zunahme der Fehlervarianzen mit wachsenden $\hat{y}_i$ bzw. x_i. Wie in Abschnitt 10.2 erläutert, lassen inhomogene Varianzen den Erwartungswert des Regressionsschätzers unberührt, vergrößern aber seinen Standardfehler, der zudem unterschätzt wird. Wenn man es mit Stichprobendaten zu tun hat, müssen diese Konsequenzen bedacht werden.

In der Fachliteratur werden verschiedene Verfahren zur »Varianzstabilisierung« vorgeschlagen. Dazu gehören bestimmte Transformationen der Y-Variablen (siehe Norusis 1985, S. 33). Wenn, wie in unserem Beispiel die Fehlervarianzen trichterförmig mit X zunehmen, ist z. B. eine logarithmische Transformation angezeigt (siehe Schlittgen 1987, S.387):

$$(10\text{-}32) \qquad \log(Y) = a + bX + e \Rightarrow E(Y) = e^{a+bx}$$

Eine solche Transformation dürfte aber, obwohl sie tatsächlich zu einer Varianzstabilisierung führt, aus inhaltlichen Gründen wenig sinnvoll sein, denn das Modell (10-32) impliziert (bei positivem b) die Annahme, daß der SPD-Stimmenanteil exponentiell mit dem Industrialisierungsgrad wächst. (Zur Interpretation von Logarithmus-Funktionen siehe Kap. 12.1)[9].

---

[9] Speziell für Anteilswerte wird auch eine Transformation mit dem Arcsinus vorgeschlagen; sie führt aber in unserem Beispiel nicht zur Varianzhomogenisierung.

Eine andere Variablen-Transformation, die für den Fall trichterförmig zunehmender Fehlervarianzen vorgeschlagen wird, läuft auf die sog. **gewichtete Regression** hinaus. Man geht von der Annahme aus, daß die Standardabweichungen der Fehler proportional mit x_i zunehmen, formal:

$$(10\text{-}33) \qquad \sigma_i = kx_i$$

wobei $k \neq 0$ eine (Proportionalitäts-)Konstante ist, deren genaue Größe wir nicht kennen. Das ist aber auch nicht nötig. Man kann zeigen, daß die Fehlervarianz unter der in (10-33) genannten Bedingung »homogenisiert« wird, wenn man die Regressionsgleichung $y_i = a + bx_i + e_i$ durch x_i dividiert:

$$(10\text{-}34) \qquad \frac{y_i}{x_i} = \frac{a}{x_i} + b\frac{x_i}{x_i} + \frac{e_i}{x_i}$$

Wir schätzen also mit der gleichen Methode wie in Abschnitt 10.1 eine neue Regressionsgleichung

$$(10\text{-}35) \qquad y_i^* = b + ax_i^* + e_i^* \quad \text{mit } y_i^* = y_i/x_i$$
$$x_i^* = 1/x_i$$
$$e_i^* = e_i/x_i$$

Die Regressionskoeffizienten sind nun vertauscht: Der Steigungskoeffizient im transformierten Modell ist der Ordinatenabschnitt des ursprünglichen Modells, und der Ordinatenabschnitt im transformierten Modell ist der Steigungskoeffizient im ursprünglichen Modell. Die zum Zwecke der Varianzhomogenisierung vorgenommene Transformation der Variablen verändert die Varianzen jedoch so, daß der Determinationskoeffizient nicht mehr aussagekräftig ist.

Wir werden die entsprechenden Regressionsergebnisse anhand unseres Beispiels gleich ermitteln, wollen zuvor aber noch zeigen, daß die Transformation geeignet ist, Varianzen anzugleichen:
Nach (10-33) ist

$$(10\text{-}36) \qquad E(e_i)^2 = (\sigma_i)^2 = k^2x_i^2$$

Nach Division durch x_i ergibt sich

(10-37)

$$E\left(\frac{e_i}{x_i}\right)^2 = \frac{1}{x_i^2} E(e_i)^2 = \frac{1}{x_i^2} \sigma_i^2 = \frac{1}{x_i^2} k^2 x_i^2 \quad \text{nach (10-33)}$$

$$= k^2 \quad , \quad \text{also eine Konstante}$$

Auch wenn man die jeweilige Größe k nicht kennt, läßt sich die Varianzhomogenisierung auf diese Weise demonstrieren. Man kann außerdem zeigen (siehe Hanushek/Jackson 1977, S. 150 ff.), daß die Transformation der beobachteten y_i- und x_i-Werte mit $1/x_i$ einer Gewichtung der einzelnen Beobachtungswerte mit dem reziproken Wert der jeweiligen Fehlervarianz gleichkommt, sofern die Annahme $\sigma_i = kx_i$ zutrifft. Beobachtungswerte, die aus einer (bedingten) Verteilung mit relativ hoher Streuung stammen, erhalten somit durch die Transformation implizit ein niedrigeres Gewicht als Beobachtungswerte, die aus einer Verteilung mit geringerer Streuung kommen. Wenn man annehmen will, daß nicht die Standardabweichung, sondern die Varianz proportional zu X ansteigt, muß die Transformation nicht mit $1/x_i$, sondern mit $1/\sqrt{x_i}$ vorgenommen werden (ebd., S. 158 f.). Natürlich lassen sich noch andere Muster inhomogener Fehlervarianzen denken, die zu weiteren Transformations- bzw. Gewichtungsformen führen. Dieses als **gewichtete Regression** bezeichnete Verfahren (**WLS**, Weighted Least Squares), ist eine Spezialform der verallgemeinerten Kleinstquadratmethode (**GLS**, Generalized Least Squares), mit der auch Korrelationen (falls bekannt) der Fehler untereinander berücksichtigt werden können (siehe Hanushek/Jackson 1977, Kap. 6).

Die OLS-Schätzung der Parameter des transformierten Modells (10-35), d.h. die indirekte WLS-Schätzung führt in unserem Analysebeispiel zu folgendem Ergebnis:

(10-38)

$$\dot{Y}^* = 0{,}934 - 10{,}439X^* \qquad \text{für die transformierten Werte}$$

$$\dot{Y} = -10{,}439 + 0{,}934X \qquad \text{für die Originalwerte}$$

Die Neuschätzung (mit WLS) korrigiert also in erheblichem Maße die ursprüngliche Schätzung (mit OLS), die einen Steigungskoeffizienten von $b = 0{,}71$ und einen Ordinatenabschnitt von $a = -1{,}42$ ergeben hatte. Die Regressionsgerade verläuft nun erheblich steiler. Dagegen ist der Standardfehler mit $s_b = 0{,}039$ etwas niedriger als zuvor (0,047). Das kommt nicht unerwartet, da wir schon in Abschn. 10.2 festgestellt hatten, daß

nicht-konstante Varianzen in der OLS-Regression zu einer Aufblähung der Standardfehler führen. (Da sie allerdings mit der ungewichteten Regression unterschätzt werden, kann es vorkommen, daß mit der WLS-Methode keine niedrigeren Standardfehler geschätzt werden als mit der OLS-Methode.)

Im Verhältnis zum Standardfehler ist die Differenz der OLS- und der WLS-Schätzer für den Steigungskoeffizienten in diesem Beispiel außerordentlich hoch, vor allem, wenn man bedenkt, daß beide Schätzer theoretisch den gleichen Erwartungswert haben. Wir müssen also annehmen, daß die von uns vorgenommene Transformation der Variablen nicht angemessen war. Die Zunahme des Steigungskoeffizienten resultiert wohl daraus, daß die hoch industrialisierten Wahlkreise, die mit ihren SPD-Stimmenanteilen relativ weit nach unten streuen, nun ein relativ geringes Gewicht erhalten (siehe Abb. 10.2 und 10.6). Wie wir schon angedeutet haben, lassen theoretische Überlegungen die in Gleichung (10-33) formulierte Annahme über die Struktur der Fehlervarianzen als nicht korrekt erscheinen. Eher ist es so, daß die Form der Beziehung (als linear) falsch spezifiziert ist und in dem bivariaten Modell relevante Regressorvariablen fehlen. Dies führt zu empirischen Fehlervarianzen, die von den »wahren« Fehlervarianzen zu stark abweichen. Bevor man auf Grund der **empirischen** Fehlervarianzen eine varianzstabilisierende Transformation bzw. eine Fallgewichtung wählt, sollte man also sicher sein, daß das Modell korrekt spezifiziert ist. Die gewichtete Regression ist kein Heilmittel gegen Fehlspezifikation. (Eine starke Differenz zwischen OLS- und WLS-Schätzungen kann aber auf eine Fehlspezifikation hinweisen.)

Eine andere Situation liegt vor, wenn man inhomogene Fehlervarianzen nicht empirisch feststellt, sondern **theoretisch** ableiten kann. Im Falle von Stichproben-Anteilswerten p_i wissen wir z. B., daß ihre Varianzen von der Zahl der Fälle n_i, über die aggregiert wird, und von der Größe der wahren Anteilswerte (π_i) abhängen (siehe Gleichung 8-42):

$$(10\text{-}39) \qquad \sigma_{p(i)} = \sqrt{\frac{\pi_i(1-\pi_i)}{n_i}} \quad , \; p_i = \pi_i$$

Die Fallzahlen n_i ergeben sich in unserem Beispiel aus der Zahl der Wahlberechtigten in den einzelnen Wahlkreisen, allgemein aus dem Umfang der Stichprobe, aus der die Anteilsgrößen berechnet wurden. Daraus folgt die in der Literatur häufig gegebene Empfehlung, bei Regressionen mit Anteilsgrößen alle Variablen mit dem reziproken Ausdruck von (10-39) zu gewichten:

$$(10\text{-}40) \quad \text{GEWICHT} = \sqrt{\frac{n_i}{\pi_i(1-\pi_i)}}$$

Wenn wir unsere ursprüngliche Gewichtung entsprechend der Größe der Wahlkreise durch die Gewichtung gemäß (10-37) ersetzen, erhalten wir folgendes Ergebnis

a= - 9.66 b= 0.80

Unbefriedigend an diesem Ergebnis ist der stark negativ geschätzte Ordinatenabschnitt, der die tatsächlichen SPD-Stimmenanteile unterschätzen läßt. Der Mittelwert der SPD-Stimmenanteile liegt nach der Gewichtung nur noch bei knapp 19 %. Auch dieses Gewichtungsverfahren ist also in Verbindung mit der linearen Regression problematisch, wenn tatsächlich eine andere Beziehungsform vorliegt. Neben der Erweiterung des Modells um zusätzliche Variablen (siehe Kap. 11) bleiben uns noch Formen der nicht-linearen Regression, die wir in den Kap. 12 besprechen.

Kehren wir nun zu den anderen Elementen der Residuenanalyse zurück. Zu ihnen gehört die Überprüfung der Normalverteilungsannahme. Das Histogramm der standardisierten Residuen (s. Abb. 10.9) läßt erkennen, daß die Abweichungen vom Modell der Normalverteilung nicht allzu groß sind.

Dieser Eindruck wird durch den »Normal Probability Plot« bestätigt (s. Abb. 10.10). In ihm wird die kumulierte relative Häufigkeitsverteilung der Residuen den kumulierten relativen Häufigkeiten der Standardnormalverteilung gegenübergestellt. (In Teil I, Abschn. 3.2 haben wir diesen Plot als »QQ-Diagramm« bezeichnet.) Wenn die Punkte alle exakt auf der Winkelhalbierenden lägen, würde das eine perfekte Anpassung des Normalverteilungsmodells belegen.

Wir hatten aber schon weiter oben festgestellt, daß dies für die Schätzung der Regressionsparameter und für den t-Test bei Fallzahlen von $n > 30$ ziemlich irrelevant ist.

Eine weitere Funktion der Residuenanalyse ist die Identifikation von Ausreißern, von Fällen also, deren Y-Werte besonders weit von der Regressionsgerade entfernt liegen. Übliche Kriterien sind Entfernungen von über zwei oder drei Standardabweichungen. In kleinen Stichproben können einzelne Ausreißer die Regressionsgerade in hohem Maße bestimmen und unter Umständen ein ziemlich irreführendes Gesamtbild vermitteln. Deshalb schätzt man die Regressionsgerade in solchen Fällen ein zweites Mal, nachdem man zuvor die Ausreißer eliminiert hat. (Allerdings

sollte man beide Ergebnisse präsentieren). Oder man wendet die bereits erwähnten, hier aber nicht erläuterten »robusten« Schätzverfahren an (siehe z. B. Hoaglin et al. 1985). Auch in großen Stichproben wie der unseren ist eine Ausreißeranalyse häufig aufschlußreich, da sie auf Einflußfaktoren aufmerksam machen kann, die zunächst vielleicht unbeachtet geblieben waren. Mit den Subkommandos /RESIDUALS OUTLIERS und /CASEWISE lassen sich die Ausreißer leicht identifizieren. Die folgende Tabelle stellt die Fälle mit den 10 stärksten Ausreißerwerten zusammen.

WKNR	*ZRESID	Wahlkreis
275	-2.83215	Saarbrücken
276	-2.55990	Ottw. Pfalz
135	-2.46974	Beuthen-Tarnowitz
449	2.42559	Hamburg West
260	-2.36694	Gladbach
279	-2.32435	Empen Achen
136	-2.24651	Kattowitz
448	2.21361	Hamburg Ost
274	-2.19427	Saarburg-M.-S.
217	-2.17394	Wittgenstein-Siegen

Am weitesten entfernt von der Regressionsgeraden liegen mit positiven Abweichungen die (stark protestantisch dominierten) Wahlkreise Nr.449 (Hamburg-West) und Nr. 448 (Hamburg-Ost), mit negativen Abweichungen die katholisch dominierten Wahlkreise Nr.275 (Saarbrücken) , Nr.276 (Ottweiler) und Nr. 135 (Beuthen-Tarnowitz). Neben der Variablen »Industrialisierungsgrad« müssen hier also noch andere Einflußfaktoren eine erhebliche Rolle spielen. Dies könnten zum Beispiel die Verteilung der Konfessionszugehörigkeit, aber auch der in den Wahlkreisen erreichte Grad der Urbanisierung sein. Sie wären in einer multiplen Regression (siehe Kap. 11) als zusätzliche Regressoren zu berücksichtigen.

10.5 Qualitative Variablen als Regressoren

Häufig sind Einflußfaktoren nicht auf metrischem Niveau meßbar oder überhaupt definierbar. Solange aber die abhängige Variable metrisch gemessen ist, läßt sich die Regressionsanalyse, wie wir sie in den vorangegangenen Abschnitten besprochen haben, auch in diesen Fällen anwenden. Als Beispiel nehmen wir die Variable »Dominante Konfession im Wahlkreis« (NKONF12) als Regressor des SPD-Stimmenanteils. In einem Wahlkreis sind entweder die Katholiken oder die Protestanten dominant. Die Variable hat also nur zwei Ausprägungen. Sie können auf verschiedene Weise kodiert werden. Am häufigsten ist die sog. Dummy-Kodierung,

bei der eine der Ausprägungen eine Null erhält, die andere eine Eins. Es gibt noch andere Möglichkeiten der Kodierung, die z. B. in Kerlinger/Pedhazur (1973), Pindyck/Rubinfeld (1981, S. 135f) oder Rochel (1983) erläutert werden.

Wir wollen in unserem Beispiel zur Reichstagswahl von 1912 den vom Katholizismus dominierten Wahlkreisen (etwa ein Drittel aller Wahlkreise) bei der Variablen NKONF12 eine Null zuordnen, den anderen, in denen der Protestantismus stärker vertreten ist, eine Eins. Eine Variable, die nur Nullen und Einsen aufweist (»Dummy-Variable«) kann als »normaler« Regressor verwendet werden - alleine in einer bivariaten Regression oder auch zusammen mit anderen Variablen in einer multiplen Regression. In der bivariaten Regression hat die Prognosegleichung mit einem dichotomen Regressor folgendes Aussehen:

(10-41)

$\hat{y}_i = a + b \cdot 1$ für alle Fälle mit x=1 (hier: protestantisch dominierte Wahlkreise)

$\hat{y}_i = a + b \cdot 0$ für alle Fälle mit x=0 (hier: katholisch dominierte Wahlkreise)

Es gibt aber nach wie vor nur eine einzige Schätzgleichung

(10-42) $y_i = a + bx_i + e_i$

deren Parameter a und b nach dem üblichen Kleinstquadrateverfahren ermittelt werden, wobei die Regressorvariable eben nur zwei Werte, null oder eins, annimmt. Die Prognosegleichung (10-41) zeigt, wie die Parameter nun zu interpretieren sind: Der Ordinatenabschnitt gibt den Erwartungswert bzw. das arithmetische Mittel ($\bar{y} \mid x_i = 0$) derjenigen Fälle an, die mit $X = 0$ kodiert wurden, in unserem Beispiel den durchschnittlichen Stimmenanteil der SPD-Kandidaten in katholisch dominierten Wahlkreisen. Laut Ergebnisausdruck (siehe Abb. 10.11) sind das 14.9 Prozent.

Der Steigungskoeffizient b gibt an, um wieviel Y-Einheiten (hier: SPD-Prozente) sich der Mittelwert der mit $x = 1$ kodierten Fälle (protestantisch dominierte Wahlkreise) vom Mittelwert derjenigen Fälle **unterscheidet,** die mit $x = 0$ kodiert wurden (katholisch dominierte Wahlkreise). Laut Ergebnisausdruck in Abb. 10.12 betrug diese Differenz 21.8 Prozentpunkte. Das bedeutet, daß in den protestantisch dominierten Wahlkreisen die SPD-Kandidaten im Durchschnitt 14.9 % + 21.8 % = 36.7 % der Stimmen erhielten. Der Determinationskoeffizient $R^2 = 33.98$ entspricht hier dem Eta-Quadrat, das wir in Teil I, Abschn. 4.2.5 besprochen haben.

Das Verfahren läßt sich auf qualitative Regressorvariablen verallgemeinern, die k, k > 2, Kategorien enthalten. So könnte man beispielsweise eine Variable kreieren, die die Wahlkreise danach klassifiziert, welche industriellen Branchen in ihnen am stärksten entwickelt waren. Nehmen wir an, wir hätten hierzu vier Kategorien (k = 4) definiert: Bergbau, Metallindustrie, Textilindustrie, Andere. In diesem Falle müßten k - 1 = 3 Dummyvariabeln (D_1, D_2, D_3) konstruiert werden. Zunächst wählt man eine Basiskategorie, z. B. Bergbau. Die Wahlkreise, in denen der Bergbau am stärksten entwickelt ist, erhalten auf allen k-1 Dummy-Variablen den Wert Null. Für die verbleibenden Kategorien werden die Dummy-Variablen z. B. in folgender Weise definiert:

D_1 = 1 für alle Wahlkreise, in denen die Metallindustrie am stärksten vertreten ist
= 0 für alle anderen Wahlkreise

D_2 = 1 für alle Wahlkreise, in denen die Textilindustrie am stärksten vertreten ist
= 0 für alle anderen Wahlkreise

D_3 = 1 für alle Wahlkreise, in denen keine der drei spezifizierten Industrien am stärksten vertreten ist
= 0 für alle anderen Wahlkreise

Daraus ergibt sich folgendes Kodierschema:

Dominante Industriebranche	D_1	D_2	D_3
Metallindustrie	1	0	0
Textilindustrie	0	1	0
Andere	0	0	1
Bergbau	0	0	0

Mit Hilfe ihrer Wertekombinationen in den k-1 Dummyvariablen sind also alle Wahlkreise eindeutig einem der k qualitativen Merkmale zugeordnet. Die Schätzgleichung sieht nun wie folgt aus:

$$(10\text{-}43) \qquad Y = a + b_1D_1 + b_2D_2 + b_3D_3 + e$$

Die Prognosegleichungen für die Wahlkreise nehmen folgende Gestalt an:

(10-44)

$$\hat{Y} = a + b_1 \cdot 0 + b_2 \cdot 0 + b_3 \cdot 0$$

für die Wahlkreise, in denen der Bergbau am stärksten entwickelt ist

$$= a$$

$\hat{Y} = a + b_1D_1 + b_2 \cdot 0 + b_3 \cdot 0$ $= a + b_1D_1$	für diejenigen Wahlkreise, in denen die Metallindustrie am stärksten entwickelt ist
$\hat{Y} = a + b_1 \cdot 0 + b_2D_2 + b_3 \cdot 0$ $= a + b_2D_2$	für diejenigen Wahlkreise, in denen die Textilindustrie am stärksten entwickelt ist
$\hat{Y} = a + b_1 \cdot 0 + b_2 \cdot 0 + b_3D_3$ $= a + b_3D_3$	für diejenigen Wahlkreise, in denen keine der spezifizierten Industrien am stärksten entwickelt ist

Der Ordinatenabschnitt gibt also wiederum das arithmetische Mittel der Y-Variable (hier: SPD-Stimmenanteile) für diejenigen Fälle an, die das als Basis - oder Bezugskategorie definierte Merkmal aufweisen (hier: Bergbau ist die am stärksten entwickelte Industriebranche). Die (geschätzten) arithmetischen Mittel der SPD-Stimmenanteile in den anderen Wahlkreisen erhält man aus den durch die Regressionskoeffizienten angegebenen Differenzen zu den »Bergbau« - Wahlkreisen.

Diese Form der Dummy-Regression ist im Ergebnis identisch mit einem Analysedesign, das in der Literatur als »einfaktorielle Varianzanalyse« bezeichnet wird.

Abb. 10.1: Schema der experimentellen Regression

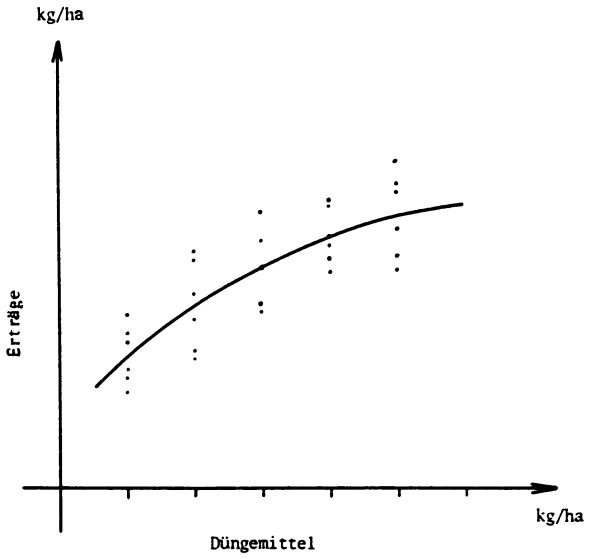


Abb. 10.2: Streudiagramm der SPD-Stimmenanteile mit Industrialisierungsgrad

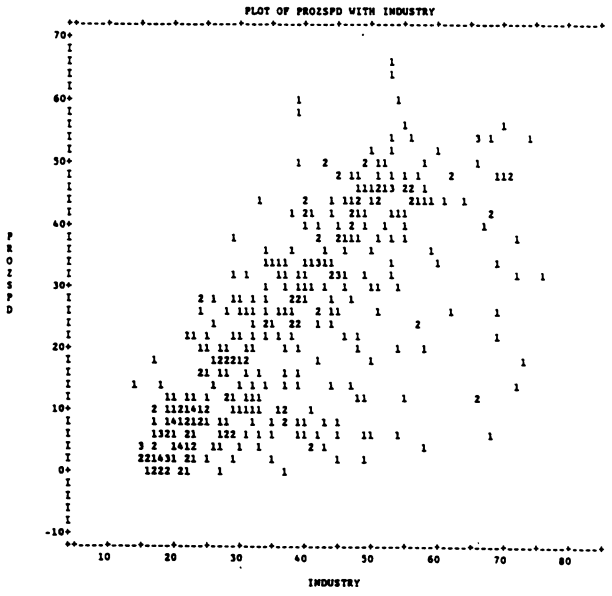


Abb. 10.3: Streudiagramm und Regressionsgerade

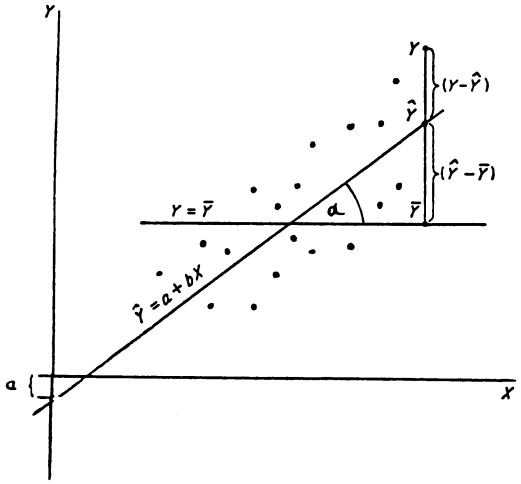
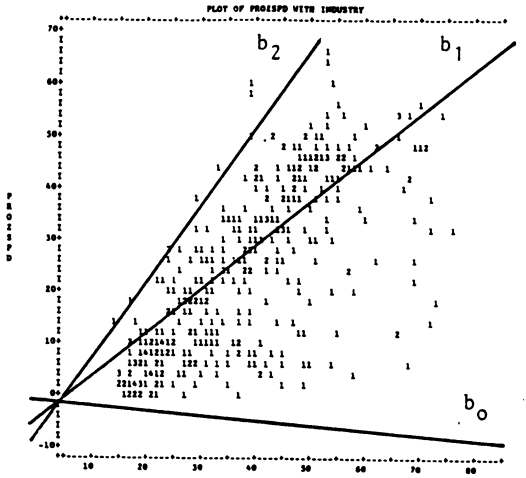


Abb. 10.4: Zur Ableitung der Regressionskoeffizienten



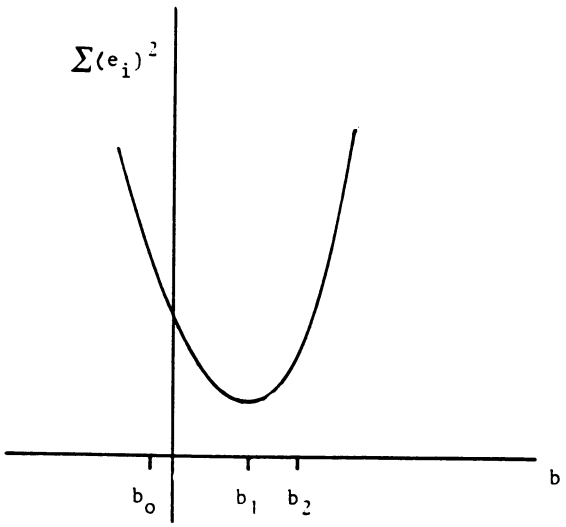
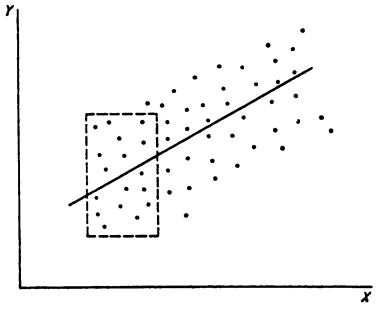


Abb. 10.5: Zum Zusammenhang von X-Varianz und Größe des Determinationskoeffizienten



Quelle: Blalock 1960, S. 291

Abb. 10.6: Streudiagramm der z-standardisierten SPD-Stimmenanteile mit z-standardisiertem Industrialisierungsgrad

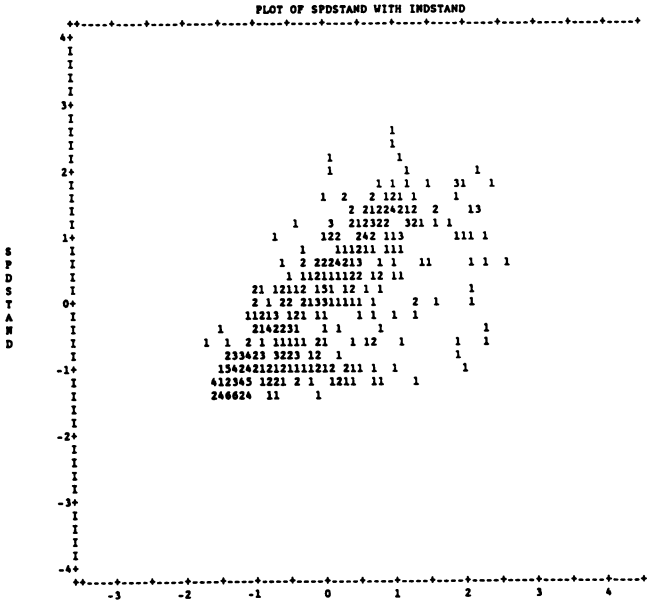


Abb. 10.7: SPSS^X-Ergebnisausdruck zur bivariaten Regression der SPD-Stimmenanteile auf den Industrialisierungsgrad des Wahlkreises (Auszug)

Multiple R	.61113
R Square	.37348
Adjusted R Square	.37189
Standard Error	14.03433

Analysis of Variance

	DF	Sum of Squares	Mean Squa
Regression	1	46143.90521	46143.905
Residual	393	77406.27369	196.962

F = 234.27758 Signif F = .0000

Variables in the Equation

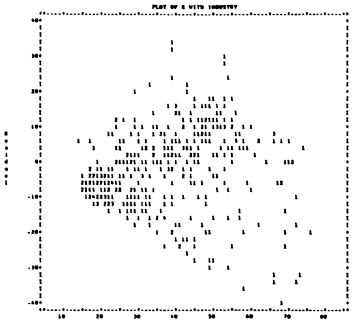
Variable	B	SE B	95% Confdnce Intrvl B		Beta	T	Sig T
INDUSTRY	.713316	.046603	.621693	.804939	.611133	15.306	.0000
(Constant)	-1.422603	2.138623	-1.422603	-1.422603		-.665	.5063

Symbols:

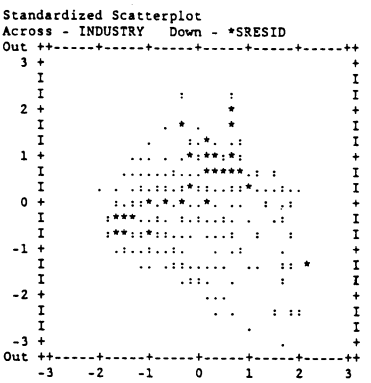
	Max N
.	2.0
:	4.0
*	10.5

Abb. 10.8: Verschiedene Residuenplots

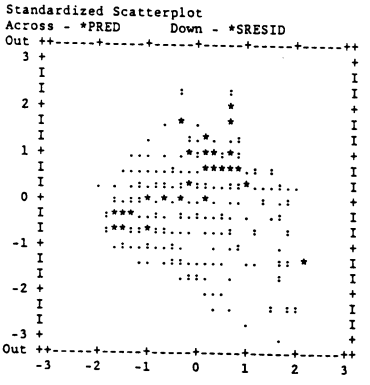
a) nicht-standardisierte Residuen mit Industrialisierungsgrad



b) s-standardisierte Residuen mit Industrialisierungsgrad



c) s-standardisierte Residuen mit SPD-Prognosewerten



d) z-standardisierte Residuen mit SPD-Prognosewerten

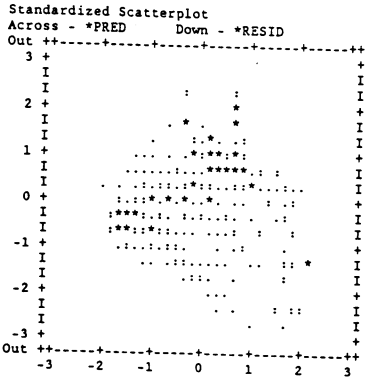


Abb. 10.9: Histogramm der s-standardisierten Residuen

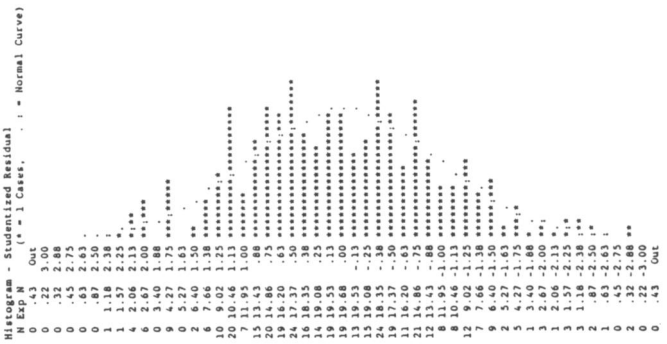


Abb. 10.10: QQ-Diagramm der Residuen mit der Standardnormalverteilung

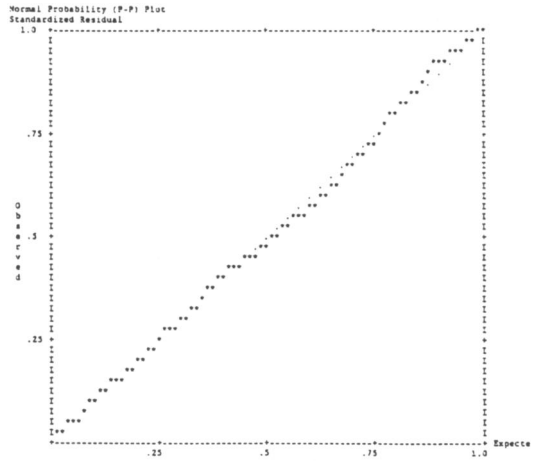


Abb. 10.11: Ergebnisausdruck zur Dummy-Regression - SPD-Stimmenanteil in Abhängigkeit von dominanter Konfession (Basiskategorie: katholisch)

EQUATION NUMBER 1 DEPENDENT VARIABLE.. PROZSPD

Variable(s) Entered on Step Number 1.. NKONF12

Multiple R	.58290	Analysis of Variance			
R Square	.33977		DF	Sum of Squares	Mean Square
Adjusted R Square	.33809	Regression	1	42059.88907	42059.88907
Standard Error	14.39087	Residual	395	81730.27797	207.09709
		F = 203.09261		Signif F = .0000	

---------- Variables in the Equation ----------

VARIABLE	B	SE B	BETA	T	SIG T
NKONF12	21.821907	1.531248	.582896	14.251	.0000
(CONSTANT)	14.908558	1.248969		11.937	.0000

KAPITEL 11

Multiple Regression

11.1 Linear-additive Beziehungen

Bisher haben wir (in Kap. 10) die SPD-Stimmenanteile in den einzelnen Wahlkreisen nur in Abhängigkeit vom Grad der Industrialisierung betrachtet. Damit sind sicherlich nicht alle relevanten Einflußgrößen erfaßt. Wir erweitern das Modell zunächst um eine Konfessionsvariable: den Anteil der Protestanten an der Bevölkerung des Wahlkreises, abgekürzt EV12. Wir gehen davon aus, daß die SPD in Wahlkreisen mit hohem Protestantenanteil (d.h. mit niedrigem Katholikenanteil) größere Stimmenanteile erreicht, als in Wahlbezirken mit niedrigem Anteil an Protestanten. Wir unterstellen (zunächst), daß Konfessions - und Industrialisierungsvariable additiv, nicht interaktiv (multiplikativ, siehe Abschn. 11.3) den SPD-Stimmenanteil beeinflussen (siehe Abb. 11.1; vgl. Teil I, Kap. 5.2.1). Formal läßt sich dieses theoretische Modell so ausdrücken:

$$(11\text{-}1) \qquad Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

und als Schätzgleichung auf unsere Variablen bezogen:

$$(11\text{-}1') \qquad \text{SPDPROZ} = a + b_1(\text{INDUSTRY}) + b_2(\text{EV12}) + e$$

Statt α und a werden oft auch die Buchstaben β_0 und b_0 eingesetzt. Die Voraussetzungen, unter denen wir die Regressionskoeffizienten und ihre Standardabweichungen schätzen (wir gehen wieder davon aus, daß unsere Beobachtungen Stichprobendaten darstellen), sind die gleichen, die wir schon für das bivariate Modell eingeführt haben (siehe Abschn. 10.2). Die dort unter Ziffer 8 gemachte Annahme, daß zwischen den Regressorvariablen keine perfekten linearen Zusammenhänge bestehen, wird natürlich erst jetzt relevant.

Die Regressionskoeffizienten a, b_1 und b_2 werden wieder nach dem Kleinstquadratverfahren berechnet. Die Koeffizienten werden also so gewählt, daß die Summe der Fehlerquadrate ein Minimum darstellt. Mathematisch läßt sich das wiederum mit Hilfe der Differentialrechnung durchführen. Die Gleichung

$$(11\text{-}2) \qquad \Sigma_i e_i^2 = \Sigma[y_i - (a + b_1 X_{1i} + b_2 X_{2i})]^2$$

wird partiell differenziert, zunächst im Hinblick auf das Absolutglied »a«, dann nacheinander hinsichtlich der Steigungskoeffizienten b_1 und b_2. Analog zu den Gleichungen (10-7) erhält man dabei einen Satz von K+1 »Normalgleichungen«, wobei K die Zahl der unabhängigen Variablen angibt (hier K=2). Wir verzichten auf Ableitung und Wiedergabe dieser Normalgleichungen, da sie im Prinzip gegenüber dem bivariaten Modell nichts Neues darstellen und in der Praxis auch nicht »per Hand« abgeleitet bzw. gelöst werden müssen.

Solange wir es nur mit zwei Regressorvariablen zu tun haben, läßt sich das Modell auch geometrisch veranschaulichen (siehe Abb. 11.2). Jeder untersuchte Fall i (i=1,2, ..., n), für den Beobachtungsdaten vorliegen, wird im dreidimensionalen Raum durch einen Punkt repräsentiert, in dem sich die Koordinaten seiner Y-, X_1- und X_2-Werte schneiden. In diesen Raum kann man sich eine Fläche (oder Ebene) eingezogen denken, die einen bestimmten Neigungswinkel zur X_1-Achse und einen bestimmten Neigungswinkel zur X_2-Achse aufweist. Die Fläche ist so zu positionieren, daß die einzelnen Punkte »minimal« von ihr abweichen (»minimal« im Sinne der Summe ihrer Abweichungsquadrate). Diese Veranschaulichung versagt natürlich, sobald das Regressionsmodell mehr als zwei unabhängige Variablen enthält. Die »Hyperebene« ist dann nur noch mathematisch abzuleiten, aber optisch nicht mehr darstellbar.

Wie sind nun die Koeffizienten eines multiplen Regressionsmodells zu interpretieren? Die Schätzer für unser Beispielmodell (11-1') sind dem Ergebnisausdruck in Abb. 11.3 zu entnehmen.

Der Koeffizient a= - 15.72, der Odinatenabschnitt, bleibt weiterhin eine rechnerische Größe, die nichts anderes darstellt, als den (hier sehr unrealistischen) Erwartungswert für Y unter der Bedingung, daß $X_1 = X_2 = 0$ ist. Spannender ist die Interpretation der beiden Steigungskoeffizienten b_1 und b_2. Der Koeffizient $b_1 = 0.645$ gibt an, um wieviel Einheiten sich die abhängige Variable (SPD-Stimmenanteil) verändert, wenn X_1 (der Industrialisierungsgrad) um eine Einheit zunimmt, während die zweite Regressorvariable, hier der Protestantenanteil, »konstant« bleibt. Entsprechend gibt der Koeffizient $b_2 = 0.273$ an, um wieviel Einheiten sich die abhängige Variable verändert, wenn X_2 bei Konstanthalten von X_1 zunimmt. Wenn der Protestantenanteil um 10 % ansteigt, wächst der SPD-Stimmenanteil im Durchschnitt um 2.7 Prozentpunkte. Oder, anders ausgedrückt: wenn der Anteil der Nicht-Protestanten (im wesentlichen also der Katholiken) um 10% zunimmt, nimmt der SPD-Stimmenanteil um 2.7 Prozentpunkte ab. Falls noch mehr Variablen in das Modell einbezogen sind, gelten entsprechende Interpretationen: Jeder Koeffizient b_k

($k = 1,2, \ldots, K$) ist ein **partieller** Koeffizient, für den mathematisch die Bedingung »simuliert« wird, daß alle anderen Regressorvariablen (neben X_k) konstant gehalten werden.

Mit dem partiellen Differenzieren der Gleichung (11-1) wird also etwas ähnliches erreicht wie durch die Drittvariablenkontrolle in der Tabellenanalyse (siehe Teil I, Kap. 5). Um den Einfluß einer Testvariable T auf die Korrelation zweier anderer Variablen X und Y untersuchen zu können, mußte dort jedoch die Stichprobe geteilt werden: für jede Kategorie der Testvariable mußten die bivariaten Verteilungen von X und Y in Partialtabellen dargestellt werden. Auch bei relativ großen Stichproben konnte man schon bei zwei oder mehr Testvariablen in die unerfreuliche Situation geraten, nicht alle Zellen der Partialtabellen mit einer ausreichend großen Zahl von Fällen besetzen zu können. Der große Vorteil des Regressionsmosdells liegt darin, daß es **partielle** Effekte, die **spezifischen** Einflüsse einer Variablen X_k auf eine andere Variable Y feststellen läßt, ohne daß man die Stichprobe in Subgruppen einteilen muß. Das »Konstanthalten« aller anderen Einflußfaktoren (soweit sie im Modell berücksichtigt sind) geschieht simultan allein mittels bestimmter algebraischer Operationen. Auf diese Weise wird es möglich, eine große Zahl von K Regressorvariablen gleichzeitig zu berücksichtigen, wobei für jeden einzelnen Regressor X_k ($k = 1,2, \ldots, K$) die übrigen K - 1 Regresoren $\#X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K\}$ als Kontrollvariablen fungieren. Das heißt, alle Steigungskoeffizienten, die wir ermitteln, schätzen den spezifischen Einfluß, den eine unabhängige Variable X_k auf eine abhängige Variable Y ausübt.

Diese Kontrolleffekte lassen sich in unserem Beispiel mit zwei Regressorvariablen wie folgt veranschaulichen: Wir führen zunächst eine Regression der SPD-Stimmenanteile auf den Protestantenanteil durch:

```
(11-3)      SPDPROZ = a + b(EV12) + e
```

Die Residuen $\{e_i\}$ speichern wir unter dem Namen RESSPD. Als nächstes regredieren wir den Industrialisierungsgrad auf den Protestantenanteil:

```
(11-4)      INDUSTRY = a + b(EV12) + e
```

Die neuen Residuen speichern wir unter dem Namen RESIND. Beide Residuenvariablen, RESSPD und RESIND, enthalten keine Varianzanteile, die durch die EV12 erklärt werden könnten (zwischen Residual- und Regressorvariable besteht vorraussetzungsgemäß keine Korrelation). Man spricht auch davon, daß die Variable »Protestantenanteil« aus den Variablen »SPD-Stimmenanteil« und »Industrialisierungsgrad« **auspartialisiert** worden ist. Wenn wir nun in einem dritten Schritt auch die **Residuen** des SPD-Stimmenanteils auf die **Residuen** der Industrialisierungsvariable regredieren,

(11-5) RESSPD = a + b(RESIND) + e

erhalten wir einen Steigungskoeffizienten b = 0.645, der mit dem identisch ist, den wir im ersten Rechengang für Gleichung (11-1') ermittelt hatten. Folglich muß es sich bei b_1 um einen **Partialkoeffizienten** handeln, der den **spezifischen** Effekt des Industrialisierungsgrades auf den SPD-Stimmenanteil wiedergibt, nachdem der Einfluß des Protestantenanteils »eliminiert« worden ist[1]. Entsprechendes gilt nun auch hinsichtlich des Steigungskoeffizienten b_2 in Gleichung (11-1'). Auch er ist ein Partialkoeffizient; er stellt den spezifischen Einfluß der Konfessionsvariable auf den SPD-Stimmenanteil dar, nachdem der Effekt der Industrialisierungsvariable »ausgeschaltet« worden ist.

In unserem Beispiel ist der partielle Steigungskoeffizient $b_1 = 0.64$ niedriger als der Steigungskoeffizient b = 0.71 aus der bivariaten Regression (beide Male mit der Fall-Gewichtung durch die Zahl der Wahlberechtigten). Die folgende Gleichung zeigt, daß dies nicht immer der Fall sein muß. Indizieren wir die abhängige Variable mit »1«, den ersten Regressor mit »2« und den zweiten Regressor mit »3«, so gilt

$$(11\text{-}6) \qquad b_{12.3} = \frac{b_{12} - b_{13}b_{32}}{1 - b_{23}b_{32}} = \frac{b_{12} - b_{13}b_{32}}{1 - r^2}$$

oder

$$b_{13.2} = \frac{b_{13} - b_{12}b_{23}}{1 - b_{23}b_{32}}$$

Der Parameter $b_{12.3}$ ist der partielle Steigungskoeffizient der zweiten Variable, $b_{13.2}$ derjenige der dritten Variable; die Variable, die nach dem »Punkt« im Index genannt wird, fungiert als Kontrollvariable. Steigungskoeffizienten mit nur zwei Indices (ohne Punkt) beziehen sich auf bivariate Regressionen. Die erstindizierte Variable stellt jeweils die abhängige Variable dar.

---

[1] Den gleichen Regressionskoeffizienten für den Industrialisierungsgrrad erhielten wir auch dann mit der Gleichung (11-1'), wenn zuvor die Konfessionsvariable lediglich aus dem Industrialisierungsgrad gemäß Gleichung (11-4), nicht aber aus der abhängigen Variablen auspartialisiert worden wäre. Man spricht in diesem Falle von einem »semipartiellen Koeffizienten«. Während partieller und semipartieller Steigungskoeffizient numerisch identisch sind, trifft dies bezüglich des Korrelationskoeffizienten nicht zu (siehe unten). Allerdings hat der semipartielle Steigungskoeffizient einen größeren Standardfehler als der partielle.

Die partiellen Steigungskoeffizienten $b_{12.3}$ bzw. $b_{13.2}$ ergeben sich aus bestimmten Kombinationen und Relationen der bivariaten Steigungskoeffizienten, wobei neben dem absoluten Betrag die Vorzeichenkombinationen eine entscheidende Rolle spielen. Der partielle Steigungskoeffizient, z. B. $b_{12.3}$, ist dann niedriger als der bivariate (b_{12}), wenn der Zusammenhang zwischen den beiden Regressorvariablen (2 und 3) das gleiche Vorzeichen hat, wie der Zusammenhang zwischen der abhängigen Variablen (1) und derjenigen Regressorvariable (3), die als Kontrollvariable betrachtet wird.

Gleichung (11-6) macht außerdem deutlich, daß der partielle Steigungskoeffizient mit dem bivariaten Steigungskoeffizienten identisch ist, wenn zwischen den Regressorvariablen kein Zusammenhang besteht, wenn also $b_{32} = b_{23} = 0$ ist.

Aber auch in diesem Falle empfiehlt sich eine multiple Regression an Stelle zweier bivariater Regressionen, da die Standardfehler der Steigungskoeffizienten aus der multiplen Regression niedriger sind als die entsprechenden Steigungskoeffizienten aus den bivariaten Regressionen. Diesen Sachverhalt kann man sich wie folgt klarmachen: Angenommen, das wahre Modell sei

$$(11\text{-}7) \qquad Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Statt dessen schätzen wir

$$(11\text{-}8) \qquad Y = \alpha + \beta_1 X_1 + \epsilon^* \quad , \quad \epsilon^* = \epsilon + \beta_2 X_2$$

Der Beitrag von X_2 wandert in den Fehlerterm ab, das heißt die Residualvarianz (die nicht erklärte Varianz) wird gegenüber dem wahren Modell erhöht - und damit, wie Gleichung (10-22) gezeigt hat, der Standardfehler $\sigma(_1)$ vergrößert.

11.1.1 Standardfehler und Signifikanztests

Die Standardfehler für die Steigungskoeffizienten einer multiplen Regression mit zwei unabhängigen Variablen lassen sich wie folgt berechnen:

$$(11\text{-}9) \qquad \sigma(b_1) = \sqrt{\frac{\sigma_\epsilon^2}{\Sigma(x_1 - \bar{x}_1)^2 (1 - r_{x(1)x(2)}^2)}}$$

$$\sigma(b_2) = \sqrt{\frac{\sigma_\epsilon^2}{\Sigma(x_2 - x_2)^2 (1 - r_{x(1)x(2)}^2)}}$$

Da die wahre Fehlervarianz σ_ε^2 unbekannt ist, muß sie aus den vorliegenden Daten mit s_e^2 geschätzt werden. Wir finden sie im Ergebnisausdruck (Abb. 11.3) in der Spalte MEANSQUARE und der Zeile RESIDUAL:

$$(11\text{-}10) \qquad \hat{\sigma}_\varepsilon^2 = \frac{\Sigma e_i^2}{df} = \frac{45768}{392} = 116{,}76$$

Die Zahl der Freiheitsgrade df = 392 ergibt sich aus der Zahl n der Fälle (395) minus der Zahl K + 1 der geschätzten Regressionskoeffizienten (einschließlich des Ordinatenabschnitts). Die Wurzel der Fehlervarianz ist $\hat{\sigma}_\varepsilon$ = 10.81. Diesen Betrag finden wir als STANDARD ERROR (der geschätzten Residuen) ebenfalls im Ergebnisausdruck. Die Nennergrößen der ersten Gleichung (11-9) können wir über das Subkommando DESCRIPTIVES erhalten (siehe Kap. 10): $\Sigma\ (x_1 - \bar{x}_1)^2 = 90685.30$; $r^2_{x_1x_2} = 0.0132$. Somit ist

(11-11)

$$\hat{\sigma}(b_1) = \sqrt{\frac{116{,}76}{90685{,}30 \cdot (1\text{-}0{,}0132)}} = 0{,}0361$$

Dieses Resultat finden wir im Ergebnisausdruck in der Spalte SE B (Standarderror b). Der Standardfehler hat sich gegenüber der bivariaten Regression (mit $s_{b(1)} = 0.047$) verringert, der t-Wert ist demgemäß von t = 15.31 auf t = 17.85 gestiegen. Der Spalte »Sig T« des Ergebnisausdruckes ist zu entnehmen, daß dieser Wert einem Fehlerrisiko mit mindestens vier Nullstellen nach dem Komma entspricht, also $\alpha < 0.00005$.

Wir ersparen es uns, diese Berechnungen auch für die Konfessionsvariable EV12 im einzelnen nachzuvollziehen. Für sie wird (in Spalte B) ein Steigungskoeffizient von $b_2 = 0.27$ ausgewiesen; er ist also erheblich geringer als der für die Industrialisierungsvariable. Geringer ist allerdings auch der (geschätzte) Standardfehler mit $s_{b(2)} = 0.017$. Das erklärt sich daraus, daß die Konfessionsvariable mit $s^2 = 1090.7$ eine wesentlich höhere Varianz aufweist als die Industrialisierungsvariable mit $s^2 = 230.16$. Die Neigung der Regressionsgeraden b_2 beruht, bildlich gesprochen, auf einem wesentlich breiteren »Fundament« als die der Geraden mit dem Steigungskoeffizienten b_1; die Schwankungen (von Stichprobe zu Stichprobe) sind geringer. Wegen des minimalen Standardfehlers ist auch der t-Wert (16.46) der Konfessionsvariable fast so hoch wie der t-Wert der Industrialisierungsvariable.

Erheblich gestiegen, von $R^2 = 37.3$ auf $R^2 = 63.0$ ist der Determinationskoeffizient, etwa 63 % der Variation der SPD-Stimmenanteile können allein durch Industrialisierungsgrad und Protestantenanteil im statistischen Sinne erklärt werden.

Im Ergebnisausdruck wird noch ein »Adjusted R Square« ausgewiesen, dessen Betrag etwas geringer ist als der des unkorrigierten R^2. Dieser korrigierte multiple Determinationskoeffizient wird nach folgender Formel berechnet (siehe Pindyck/Rubinfeld 1981, S. 80):

$$(11\text{-}12) \qquad R_a^2 = 1-(1-R^2)\cdot\frac{n-1}{n-(K+1)}$$

(»K« entspricht der Anzahl der unabhängigen Variablen)

Wie in Kapitel 10 erläutert, ist der Ausdruck $(1 - R^2)$ das Unbestimmtheitsmaß, der Anteil der **nicht** erklärten Variation an der Gesamtvariation. Determinationskoeffizient und Unbestimmtheitsmaß ergänzen sich also zu 1, anders ausgedrückt: $R^2 = 1-(1-R^2)$. Die Korrektur von R^2 zu R_a^2 kommt dadurch zustande, daß der negative Summand , das Unbestimmtheitsmaß, um so stärker gewichtet wird, je größer die Zahl K der Regressorvariablen, je geringer also die Zahl der Freiheitsgrade. Den Sinn dieser Operation kann man sich in etwa durch folgende Überlegung vergegenwärtigen: Wenn die Zahl der Koeffizienten (der Parameter im Regressionsmodell) ebenso groß ist wie die Zahl der Fälle, muß die Residualvarianz gleich Null sein. Bei zwei Fällen und einer Regressorvariable (neben der abhängigen Variable), geht die Regressionsgerade durch die beiden Koordinatenpunkte $(y_1;x_1)$ und $(y_2;x_2)$. Abweichungen davon treten nicht auf. Ganz gleich, ob ein struktureller Zusammenhang besteht oder nicht, wird die Residualvariation durch die Anzahl der im Modell berücksichtigten Variablen gemindert. In R_a^2 wird deshalb das Unbestimmtheitsmaß durch den Faktor $(n-1)/(n-K-1)$ entsprechend erhöht. (Weitere Erläuterungen hierzu geben Pindyck/Rubinfeld 1981, S. 78ff.).

Die statistische Signifikanz von R^2 kann mit Hilfe der F-Statistik mit K und $n-(K+1)$ Freiheitsgraden getestet werden (zur F-Verteilung siehe die Erläuterungen in Abschn. 7.4.3).

$$(11\text{-}13) \qquad F_{K,N-K-1} = \frac{R^2}{1-R^2} \cdot \frac{n-K-1}{K}$$

$$= \frac{0.63}{0.37} \cdot \frac{392}{2} = 333.1$$

Das entsprechende Fehlerrisiko bei der Zurückweisung der Nullhypothese $H_0 : R^2_{Pop} = 0$ hat laut Ergebnisausdruck (s. Abb. 10.3), mindestens vier Nullstellen nach dem Komma, $\alpha < 0.00005$.

11.1.2 Residuenplots und Modellerweiterung

Werfen wir nun noch einen Blick auf verschiedene Residuenplots, wobei wir hier ausschließlich von den s-standardisierten Residuen (*SRESID) ausgehen (siehe Abb. 11.4). Ein Vergleich von Abb. 11.4 mit Abb. 10.8 zeigt, daß die Hinzunahme des Protestantenanteils das Regressionsmodell nicht nur hinsichtlich einer größeren Prognosefähigkeit verbessert hat; auch die Residuen haben sich dem gewünschten Muster angenähert; sowohl mit dem Industrialisierungsgrad als auch mit dem Protestantenanteil bilden sie keinen ausgeprägten Zusammenhang mehr. (Es gibt ein paar »Ausreißer«, die wir hier aber nicht untersuchen wollen.) Das Streudiagramm mit den z-standardisierten Prognosewerten (*PRED) zeigt jedoch, daß von konstanter Varianz der Residuen immer noch nicht gesprochen werden kann, obwohl die Trichterform nun weniger ausgeprägt ist. Die Varianz scheint etwas oberhalb des arithmetischen Mittels ($\bar{y} = 29.590$) der SPD-Stimmenanteile besonders groß zu sein. (Wegen der Standardisierung entspricht der Wert »0« auf der X-Achse dem arithmetischen Mittel der Prognosewerte). Wie oben schon erwähnt, ist ein Varianzmaximum bei $\hat{y} = 50\%$ aus theoretischen Gründen zu erwarten (siehe auch Kap. 12.2).

Wir können natürlich nicht behaupten, daß nun alle relevanten Variablen im Modell berücksichtigt sind. Ein denkbarer Kandidat für eine weitere Einflußgröße ist der Grad der in den Wahlbezirken erreichten Urbanisierung, der Gegensatz zwischen einer stärker traditionalistisch orientierten Landbevölkerung und einer eher fortschrittlich gesonnenen Stadtbevölkerung. Ein möglicher Indikator, mit dem diese Variable angesprochen werden kann, ist der Anteil der Bevölkerung, die in Gemeinden mit über 2000 Einwohnern lebt (URBAN).
Wir schätzen somit folgendes Modell

(11-14)

$$PROZSPD = a + b_1(INDUSTRY) + b_2(EV12) + b_3(URBAN) + e$$

Das Ergebnis sieht wie folgt aus (die Standardfehler sind den Schätzern in Klammern hinzugefügt):

$a = -15.84\ (1.60)$
$b_1 = .23\ (0.047)$
$b_2 = .27\ (0.014)$
$b_3 = .30\ (0.025)$
$R^2_a = .725$

Alle Koeffizienten sind hochsignifikant. Wie erwartet, ist der SPD-Stimmenanteil um so höher, je größer der Anteil der Bevölkerung ist, der nicht auf dem Lande lebt. Gegenüber dem Modell (11-1') ist der Steigungskoeffizient des Industrialisierungsgrades von $b_1 = 0.64$ um mehr als die Hälfte auf $b_1 = 0.23$ gesunken. Ein Großteil des ihm zuvor zugeschriebenen Einflußgewichts scheint also dem Urbanisierungsfaktor zuzukommen. Beide Variablen korrelieren allerdings mit $r = 0.749$ hoch untereinander. Das macht einerseits die Trennung der beiden Einflußgewichte besonders wünschenswert, andererseits aber auch besonders riskant. Auf dieses Problem werden wir - neben anderen - im nächsten Abschnitt unter dem Stichwort »Multikollinearität« zurückkommen. Dort wird auch deutlich werden, warum der Standardfehler des Steigungskoeffizienten der Industrialisierungsvariable durch Hinzunahme des Urbanisierungsgrades wieder zugenommen hat.

11.2 Besondere Probleme des Schätzens und Testens

11.2.1 Multiples Testen

Die Nullhypothese $R^2_{Pop} = 0$ ist gleichbedeutend mit der Annahme, alle Steigungskoeffizienten des Regressionsmodells seien gleich Null; H_0: $\beta_1 = \beta_2 = \ldots = \beta_K = 0$ (»joint hypothesis«). Da drängt sich die Frage auf, ob der F-Test für den Determinationskoeffizienten nicht durch die Serie der t-Tests für alle Steigungskoeffizienten ersetzt werden kann. Es läßt sich aber zeigen, daß ein solches »multiples Testen«, das mehrfache Testen von Hypothesen am gleichen Datenmaterial, nicht identisch ist mit dem einmaligen Testen einer »joint hypothesis«. Das kann man sich mit Hilfe folgender Überlegung klar machen: Angenommen, wir legen als Signifikanzniveau fest, das Fehlerrisiko für die Zurückweisung der Nullhypo-

these ($\beta_k = 0$) solle $\alpha \leqq 0.05$ betragen. Bei 20 Versuchen, d.h. wenn das Regressionsmodell 20 erkärende Variablen enthält, die unkorreliert sind, muß man damit rechnen, daß einer der Steigungskoeffizienten, sagen wir b_k, »zufällig« in diesen »unwahrscheinlichen« Bereich $b_k \gg 0$ fällt, obwohl in der Population $\beta_k = 0$ ist. Das würde dazu führen, die Nullhypothese: alle Steigungskoeffizienten sind gleich Null, zurückzuweisen. Beim multiplen Testen ist also das tatsächliche Fehlerrisiko höher als im α - Niveau (dem Signifikanzkriterium) für den einzelnen Test angegeben.

Bei korrelierten Regressorvariablen gibt es jedoch auch einen entgegengesetzten Effekt: Korrelationen zwischen den Regressorvariablen (»Multikollinearität«) führen zu Kovarianzen zwischen den Schätzern für die Regressionskoeffizienten. Ihre Standardfehler werden um so größer, die t-Werte entsprechend niedriger, je höher die Korrelationen zwischen den unabhängigen Variablen sind. Auch dieser Effekt wird vom F-Test korrekt berücksichtigt. Bei wiederholter Anwendung des t-Tests (multiples Testen) ist nicht vorauszusehen, welcher der entgegengesetzten Effekte stärker ist. Es ist möglich, daß sich einer der Koeffizienten fälschlicherweise als signifikant erweist; ebenso ist es möglich, daß sich einer der Koeffizienten fälschlicherweise als nicht signifikant zeigt. Der F-Test über alle Steigungskoeffizienten berücksichtigt beide Tendenzen und testet die Gesamthypothese $\beta_1 = \beta_2 = \ldots = \beta_K = 0$ korrekt auf dem vorgegebenen α-Niveau. Man sollte also nicht überrascht sein, wenn bei korrelierten Regressorvariablen laut F-Test R^2 signifikant ist, während die Serie von t-Tests alle Steigungskoeffizienten bei gleichem α-Kriterium als »nicht signifikant« ausweist (siehe zu dieser Problematik auch Hays 1973, S. 478f.).

11.2.2 Multikollinearität

Zur Multikollinearität nun noch einige Hinweise:
Im Grenzfall einer perfekten Korrelation ($R^2 = 1$) zwischen zwei oder mehreren Regressorvariablen (d.h. eine Regressorvariable wird vollständig durch eine oder alle anderen Regressorvariablen determiniert) können die Regressionskoeffizienten überhaupt nicht geschätzt werden, da das Gleichungssystem in diesem Fall zwei oder mehrere Gleichungen enthält, die nicht unabhängig voneinander sind. Diese Extremsituation wird selten auftreten. Interessant ist also die Multikollinearität mittlerer Höhe. Zunächst einmal ist festzustellen, daß die Regressionsschätzer zwar erwartungstreu bleiben, daß aber ihre Effizienz um so geringer wird, je höher die Korrelation der beteiligten Regressorvariablen untereinander ist. Das bedeutet, ein spezifisches Stichprobenergenbis kann relativ weit vom wahren Wert entfernt sein. Wie bereits erwähnt, kovariieren die Schätzer zweier Steigungskoeffizienten (im Absolutbetrag) um so stärker, je stärker die entsprechenden Regressorvariablen untereinander korrelieren. In ei-

nem Modell mit zwei Regressorvariablen hat die Kovarianz der Schätzer das umgekehrte Vorzeichen wie die Korrelation der beteiligten Variablen (siehe Pindyck/Rubinfeld 1981, S. 78). Das bedeutet, daß in diesem Fall folgende Tendenz besteht: wird der eine Steigungskoeffizient überschätzt, wird der andere Steigungskoeffizient unterschätzt(ebd., S. 90). Korrelieren die beiden Regressorvariablen negativ untereinander, ist die Kovarianz zwischen den Regressionsschätzern positiv; es besteht also eine Tendenz zur gemeinsamen Über- oder Unterschätzung der beiden Steigungskoeffizienten[2].

In der Praxis kann es somit zu einer geradezu paradoxen Situation kommen: Wie oben erläutert, unterscheiden sich die Punktschätzer der Steigungskoeffizienten einer multiplen Regression von denen der entsprechenden bivariaten Regression nur, wenn die Regressorvariablen untereinander (linear) korrelieren. Man setzt also die multiple Regression ein, um den **spezifischen** Effekt einer Variablen durch »Konstanthalten« der übrigen Variablen zu ermitteln. Je höher aber ein Regressor mit den anderen Regressoren korreliert (je höher also die »Multikollinearität« ist), desto unsicherer wird die Partialisierung, desto unzuverlässiger werden die Schätzwerte für die (partiellen) Steigungskoeffizienten. Wie hoch darf die Multikollinearität sein, bevor die Ergebnisse unzuverlässig werden? Hierfür gibt es kein eindeutiges Kriterium. In der Literatur finden sich verschiedene Faustregeln, z. B. die, daß die Werte einer Regressorvariable durch die Gesamtheit der anderen Regressorvariablen höchstens mit einem multiplen Determinationskoeffizienten von $R^2 = 0.5$ vorhersagbar sein sollen. In unserem Beispielmodell (11-14) liegen wir oberhalb dieses Kriteriums, da Urbanisierung- und Industrialisierungsgrad alleine schon mit $r = 0.749$ miteinander korrelieren.

Andererseits werden die Steigungskoeffizienten aber auch verzerrt geschätzt, wenn man, um zu hohe Multikollinearität zu vermeiden, eine relevante Variable aus dem Modell eliminiert. Einen in allen Fällen überzeugenden Ausweg aus diesem Dilemma gibt es nicht, auch nicht mit den sogenannten »robusten« Regressionsverfahren und der »Ridge-Regression«, die gelegentlich empfohlen werden (siehe z. B. Opp/Schmidt 1976, S. 178)[3]. In jedem Falle ist hohe Multikollinearität eine Herausforderung an den Theoretiker: Wenn zwei Indikatoren sehr hoch untereinander korrelieren, muß er sich überlegen, ob sie nicht das gleiche Phänomen, den gleichen Prozeß messen; zumindest muß er erwägen, ob die einzelnen

---

[2] In einem Modell mit mehreren Regressorvariablen müßten jeweils die Korrelationsbeziehungen mit allen anderen Variablen berücksichtigt werden, so daß die Gesamteffekte nur nach relativ komplexen Berechnungen angegeben werden können (siehe Belsley et al. 1980).

[3] Ein aufwendiges Verfahren, Ausmaß und Folgen der Multikollinearität zu analysieren, schlagen Belsley/Kuh/Welsch (1980) vor.

Variablen nicht als Subdimensionen einer abstrakteren theoretischen Kategorie interpretiert werden können. Urbanisierung und Industrialisierung können für lange historische Zeiträume nicht als einheitlicher Prozeß aufgefaßt werden, für die europäischen Länder ab 1850 vielleicht doch. Soziologen haben z. B. die beiden Prozesse, von denen hier die Rede ist, unter der Kategorie der »Modernisierung« zusammengefaßt.

Wir können den theoretischen Nutzen solcher Abstraktionen hier nicht erörtern. Es sei aber darauf hingewiesen, daß inzwischen Regressionsverfahren entwickelt worden sind (z. B. sog. LISREL-Modelle), mit denen man mehrere Variablen simultan als Indikatoren eines (einzigen) theoretischen Konstrukts berücksichtigen kann (zur Erläuterung siehe z. B. Pfeifer/Schmidt 1987).

11.2.3 Weitere Aspekte des F-Tests (*)

Kehren wir noch einmal zum F-Test zurück. Um die Prüfgröße (11-13) in einer bestimmten Weise umformen zu können, führen wir folgende Abkürzungen ein:

$SAQ_G = \Sigma (y_i - \bar{y})^2$ - für die Gesamtsumme aller Abweichungsquadrate
$SAQ_M = \Sigma (\hat{y}_i - \bar{y})^2$ - für die durch die Regressionsvariable(n), das Modell, erklärte Summe der Abweichungsquadrate
$SAQ_R = \Sigma (y_i - \hat{y}_i)^2$ - für die nicht erklärte Restsumme der Abweichungsquadrate

Somit kann man die Prüfgröße »F« gemäß (11-13) wie folgt schreiben (vgl. Kap. 10 die Gleichungen (10-13)ff.)

(11-15)

$$F_{K,n-K-1} = \frac{R^2}{1-R^2} \cdot \frac{n-K-1}{K}$$

$$= \frac{SAQ_M/SAQ_G}{SAQ_R/SAQ_G} \cdot \frac{n-K-1}{K}$$

$$= \frac{SAQ_M}{SAQ_R} \cdot \frac{n-K-1}{K}$$

$$= \frac{SAQ_M/K}{SAQ_R/n-K-1} = \frac{77782/2}{45768/392} = \frac{38891}{117} = 333$$

Auf diese Weise läßt sich der F-Wert so rekonstruieren, wie er laut Ergebnisausdruck in Abb. 11-3 als »Analysis of Variance« konzipiert worden ist. Im Zähler wie im Nenner stehen Varianzschätzer, denn Varianzschätzer sind, wie wir schon früher festgestellt haben, allgemein definiert als »Variation dividiert durch die Zahl der Freiheitsgrade.«[4] Die nicht erklärte Variation SAQ_R wird zwar aus n Stichprobenfällen berechnet; doch haben wir mit diesen Beobachtungsdaten schon die drei Parameter des Regressionsmodells (K=2 Steigungskoeffizienten plus einen Ordinatenabschnitt) geschätzt. Für die Schätzung der Populationsvarianz stehen also nur noch n - 3 unabhängige Informationen zur Verfügung; die »letzten« drei Werte y_n, y_{n-1}, y_{n-2} liegen fest, wenn die 3 Regressionsparameter und die Werte $y_1, y_2, \ldots, y_{n-3}$ gegeben sind.

Die K Freiheitsgrade für den Zählerausdruck ergeben sich rechnerisch aus einer Differenz: $SAQ_M = SAQ_G - SAQ_R$; die Gesamtvariation läßt sich, wie wir in Kap. 10 sahen, unter bestimmten Bedingungen in zwei unabhängige Komponenten, SAQ_M und SAQ_R, zerlegen. Der SAQ_G sind n-1, der SAQ_R n-1-K Freiheitsgrade zugeordnet, somit stehen für die Differenz SAQ_M noch (n-1)-(n-1-K)=K Freiheitsgrade zur Verfügung[5]. Die letzte Zeile in Gleichung (11-15) kann also wie folgt geschrieben werden:

$$(11\text{-}15') \qquad F_{K,n-K-1} = \frac{(SAQ_G - SAQ_R)/[(n-1)-(n-1-K)]}{SAQ_R/n-K-1}$$

Die Differenzenbildung im Zähler des F-Quotienten läßt sich inhaltlich wie folgt deuten: SAQ_G stellt die Residuen aus einem extrem eingeschränkten Modell dar, das nur einen zu schätzenden Parameter enthält, nämlich den Ordinatenabschnitt, der mit dem arithmetischen Mittel $\bar{y}$ identisch ist

$$(11\text{-}16) \qquad y_i = a + e_i \qquad ; \quad a = \bar{y}$$

$$\hat{y}_i = \bar{y}$$

Die SAQ_R dieses »eingeschränkten« Modells, die wir jetzt als SAQ_{Re} be-

---

[4] Daß es sich bei beiden um unabhängige Schätzer für die Populationsvarianz σ_y^2 handelt, ist intuitiv kaum nachvollziehbar und wird hier auch nicht erläutert.

[5] Dabei stützt man sich auf ein statistisches Theorem über die Zerlegbarkeit sog. »quadratischer Formen«, das hier nicht erläutert werden kann.

zeichnen wollen, werden verglichen mit den SAQ_{Rv} eines »vollständigen« (bzw. erweiterten) Modells, z. B.

$$(11\text{-}17) \qquad y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$$

Es wird also gefragt, um welchen Betrag die SAQ_{Re} zusätzlich durch die eingeführten Modellparameter (hier b_1 und b_2) auf die SAQ_{Rv} »reduziert« wird. Ein solcher Vergleich kann immer dann vorgenommen werden, wenn die Parameter des »eingeschränkten« Modells eine Teilmenge der Parameter des »vollständigen« Modells darstellen. Die Attribute »eingeschränkt« oder »vollständig« sind also relativ gemeint. Das Modell (11-17) ist z. B. ein eingeschränktes Modell gegenüber dem »vollständigen« Modell

$$(11\text{-}18) \qquad y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i}$$

sofern die X_1- und X_2-Variablen in beiden Modellen dieselben sind. In unserem Beispiel sind X_1 die Industrialisierungs-, X_2 die Konfessionsvariable und X_3 der Urbanisierungsgrad.

Den Ergebnisausdruck zum Modell (11-17) bzw. (11-1') haben wir schon im vorigen Abschnitt ausführlich besprochen (siehe Abb. 11.3). Den Ergebnisausdruck zum erweiterten Modell (11-18) finden wir in Abb. 11.5. Beide Modelle lassen sich in SPSS^x mit einem einzigen REGRESSION-Kommando und den entsprechenden Subkommandos schätzen:

```
REGRESSION VARIABLES=PROZSPD,INDUSTRY,EV12,URBAN
  /STATISTICS=DEFAULT,CHA,BCOV
  /DESCRIPTIVES=DEFAULTS
  /DEPENDENT=PROZSPD
  /ENTER=EV12, INDUSTRY/ENTER=URBAN
```

Man beachte, daß in der Variablenliste alle Variablen genannt werden, die im vollständigen Modell enthalten sind. Durch das Subkommando ENTER lassen sich die Regressoren aus diesem Variablensatz spezifizieren: mit dem ersten ENTER-Kommando werden die Variablen des eingeschränkten Modells, mit dem zweiten ENTER-Befehl die **zusätzlichen** Variablen des erweiterten Modells eingeführt. Mit dem STATISTICS-Kommando läßt sich unter anderem der Parameter CHA anfordern. Er liefert die F-Statistik für die zusätzlichen Parameter, für die durch sie zusätzlich erklärten Varianzanteile.

Für das erweiterte Modell wird **insgesamt** ein F-Wert von 348.05 mit einem $\alpha < 0.0005$ ausgewiesen. Was uns aber primär interessiert, ist die Frage, ob die neu in das Modell aufgenommene Variable »Urbanisierungsgrad« einen **zusätzlichen** Anteil der Variation der Y-Werte erklärt - über

die Variationsanteile hinaus, die schon durch den Industrialisierungsgrad und den Protestantenanteil erklärt worden sind. Wir müssen also die Differenz der SAQ's aus den beiden Modellen betrachten:

Die Summe der Abweichungsquadrate beträgt im eingeschränkten Modell (11-17) $SAQ_{Re} = 45768$ mit 392 Freiheitsgraden, im vollständigen Modell (11-18) $SAQ_{Rv} = 33660$ mit 391 Freiheitsgraden. Um zu testen, ob die Differenz SAQ_{Re}-SAQ_{Rv} signifikant ist, berechnen wir die Prüfgröße F gemäß (11-15'):

$$(11\text{-}19) \qquad F_{1;391} = \frac{(45768-37442)/(392-391)}{37442/391} = 140{,}64$$

Diesen Wert finden wir im 2. Teil des Ergebnisausdrucks (Abb. 11.5) neben der Bezeichnung »F-Change«, darunter die Information, daß dieser F-Wert hochsignifikant ist. Die Nullhypothese, daß die dritte Variable (Urbanisierungsgrad) keine **zusätzliche** Varianz erklärt, kann mit einem Fehlerrisiko von $\alpha < 0.00005$ zurückgewiesen werden.

Wenn sich das vollständige bzw. erweiterte Modell von dem eingeschränkten nur durch **eine** zusätzliche Regressorvariable unterscheidet, führt der F-Test gemäß (11-19) zum gleichen Ergebnis wie der t-Test für die entsprechende Variable. Wir hatten ja festgestellt, daß die Steigungskoeffizienten **partielle** Einflußgewichte darstellen. Sie sind ungleich Null nur, wenn die jeweilige Regressorvariable einen Anteil der Y-Variable erklärt, der nicht durch die anderen Regressorvariablen, die sich im Modell befinden, erklärt werden kann. Durch algebraische Operationen »simuliert« der t-test für jede unabhängige Variable eine Situation wie sie im »F-Change-Test« vorliegt. Schon in Abschnitt 7.4.3 hatten wir auf die Beziehung $\sqrt{F_{1,n}} = t_n$ hingewiesen. In der Tat führt die Wurzel des F-Wertes aus (11-19) zu dem t-Wert, der im Ergebnisausdruck (Abb. 11.5) für URBAN ausgewiesen ist:

$$(11\text{-}20) \qquad \sqrt{F_{1,391}} = t_{391}$$

$$\sqrt{140{,}64} = 11{,}86$$

Wenn das erweiterte Modell nur einen zusätzlichen Parameter enthält, können wir uns also den »F-change-test« sparen. Dieser Test wird aber immer dann interessant, wenn **Gruppen** von Variablen hinsichtlich der Signifikanz ihres Erklärungsbeitrages geprüft werden sollen. Vorausset-

zung (neben der Normalverteilung und Homoskedastizität der Residuen) ist, wie bereits erwähnt, daß die Parameter des eingeschränkten Modells eine Untermenge des erweiterten Modells sind.

11.3 Standardisierte Regressions- und partielle Korrelationskoeffizienten

Die Steigungskoeffizienten, wie wir sie bisher besprochen haben, geben an, um wieviele Skaleneinheiten sich die Y-Variable ändert, wenn die jeweilige Regressorvariable um eine Skaleneinheit wächst und alle anderen Regressorvariablen konstant gehalten werden. Vergleichen wir in unserem Beispiel zwei Wahlbezirke, die hinsichtlich ihres Protestantenanteils und Urbanisierungsgrades gleich sind, sich aber im Industrialisierungsgrad um 10% unterscheiden. Das von uns geschätzte Modell (11-14) läßt erwarten, daß die SPD in dem Wahlkreis mit dem höheren Industrialisierungsgrad 2.3 Prozentpunkte mehr Stimmen erhalten hat. Wenn der Anteil der Protestanten in der Bevölkerung um 10% zunimmt, nimmt der SPD-Stimmenanteil um durchschnittlich 2.7 Prozentpunkte zu. Es liegt nahe, die Steigungskoeffizienten verschiedener Prädiktorvariablen als Indikator für die relative Einflußstärke der unabhängigen Variablen (im Kontext des gegebenen Modells) zu interpretieren. In unserem Beispiel würde man also der Konfessions-, der Industrialisierungs- und der Urbanisierungsvariable eine etwa gleich große theoretische Bedeutung zumessen, da die entsprechenden Koeffizienten dem Absolutbetrag nach etwa gleich groß sind (siehe Abb. 11.5). Hätte ein Politiker diese Ergebnisse schon 1912 oder kurz danach gekannt, hätte er möglicherweise ganz andere Ansichten über die »Bedeutung« der verschiedenen Variablen geäußert. Für ihn wären die Industrialisierungs- und die Urbanisierungsvariable vielleicht kausal wichtiger als die Konfessionsvariable. Er könnte sich nämlich ausrechnen, daß die Anteile der Protestanten an der Bevölkerung ziemlich stabil bleiben, Urbanisierung und Industrialisierung jedoch weiter fortschreiten und, in gewissem Grade, sogar gesteuert werden können. Andererseits ließe sich u.U. auch das »Image« der SPD bei den Katholiken verbessern.

In unserem Beispiel stellen alle Regressorvariablen Anteilswerte dar, Zählgrößen, die sich stets auf die gleichen Untersuchungseinheiten beziehen. Dennoch ist es nicht unproblematisch, ihre relative kausale Bedeutung gleichzusetzen mit der relativen Größe ihrer Steigungskoeffizienten. In welcher Weise sind Prozentzuwächse im Protestantenanteil und im Industrialisierungsgrad miteinander vergleichbar?[6] Das Problem der Vergleichbarkeit ist noch schwerer zu lösen, wenn die Variablen mit **unter-**

[6] Zu dieser Problematik siehe King (1986).

schiedlich konstruierten Skalen gemessen worden sind. Angenommen, man hätte den Industrialisierungsgrad in Mengen- oder Preiseinheiten industriell erzeugter Güter gemessen. Selbst wenn man diese Größe in »pro Kopf der Bevölkerung« ausgedrückt hätte, wären diese Skalenwerte nicht mehr direkt vergleichbar mit den Anteilsgrößen der anderen Variablen, die Skaleneinheiten wären unterschiedlich definiert. Wie wollte man z. B. einen Anstieg des SPD-Stimmenzuwachses pro zusätzlich erzielter Preiseinheit mit dem Stimmenzuwachs je Prozentpunkt des Protestantenanteils vergleichen?

In der Fachliteratur wird zur Lösung dieses Problems zumeist vorgeschlagen, die Regressionskoeffizienten zu sogenannten »Beta-Koeffizienten« zu standardisieren. Diese Betakoeffizienten[7] erhält man, indem man eine »normale« (multiple) Regression durchführt, zuvor aber alle Modell-Variablen (einschließlich der abhängigen Variablen) z-standardisiert. (Die z-Standardisierung haben wir schon an verschiedenen anderen Stellen besprochen, z. B. in Abschn. 10.1.1). Man kann sie (im additiven Modell) aber auch aus den nicht-standardisierten Steigungskoeffizienten berechnen:

$$(11\text{-}21) \qquad \beta_{YX} = b_{YX}\frac{s_X}{s_Y}$$

Man multipliziert den nicht-standardisierten (partiellen) Steigungskoeffizienten mit der Standardabweichung der betreffenden Regressorvariablen X und dividiert durch die Standardabweichung der abhängigen Variablen Y.

Durch die z-Standardisierung werden die Variablen sozusagen auf die Maßeinheit »Standardabweichung« geeicht. Formal gibt der Beta-Koeffizient also an, um wieviele Standardabweichungen z(y) die abhängige Variable zunimmt, wenn die Regressorvariable um eine Standardabweichung z(x) zunimmt. Im bivariaten Fall ist der standardisierte Steigungskoeffizient identisch mit dem Pearsonschen Korrelationskoeffizienten r, wie wir in Abschnitt 10.1.1 gezeigt haben.

In den SPSS^x-Ergebnisausdrucken findet man die Beta-Koeffizienten links neben den t-Werten. In Abb. 11.5 (Forts.) können wir z. B. ein Beta=.198408 für den Industrialisierungsgrad ablesen.

---

[7] Die Wahl des Buchstabens »Beta« für den Schätzer des Steigungskoeffizienten ist etwas unglücklich, da man üblicherweise auch die nicht-standardisierten »wahren« Regressionskoeffizienten mit diesem Symbol bezeichnet.

Die Beta-Koeffizienten sind als Maß für die relative Einflußstärke der verschiedenen Regressorvariablen nicht unproblematisch. Wie aus Gleichung (11-21) hervorgeht, ist die Größe von Beta abhängig vom Verhältnis der Varianzen bzw. der Standardabweichungen der abhängigen Variable und der jeweiligen Regressorvariable. Da die abhängige Variable für alle Regressorvariablen die gleiche ist, werden die Größenverhältnisse der Betas entscheidend durch die unterschiedlichen Varianzen der jeweiligen Regressorvariablen bestimmt. In unserem Regressionsmodell (11-14) z. B. haben Industrialisierungsgrad und Protestantenanteil einen etwa gleich großen unstandardisierten Steigungskoeffizienten; die Konfessionsvariable hat mit $\beta_2 = .508$ aber einen mehr als doppelt so großen Beta-Koeffizienten wie der Industrialisierungsgrad (siehe Abb. 11.5). Das erklärt sich daraus, daß die Konfessionsvariable mit $s_{x(2)} = 33.03$ eine mehr als doppelt so große Standardabweichung hat wie die Industrialisierungsvariable mit $s_{x(1)} = 15.17$.

Es wäre wohl wenig sinnvoll, anhand des Beta-Koeffizienten zu behaupten, die Konfessionsvariable habe auf den SPD-Stimmenanteil einen doppelt so großen Einfluß wie der Industrialisierungsgrad. Nehmen wir an, der unstandardisierte Regressionskoeffizient von $b_1 = .23$ drücke eine zeitlich stabile Beziehung zwischen SPD-Stimmenanteil und Industrialisierungsgrad aus. Nehmen wir gleichzeitig an, die Varianz des Industrialisierungsgrades werde über Zeit geringer, die Anteile der industriell Beschäftigten der verschiedenen Wahlbezirke glichen sich zunehmend einander an. Die Konsequenz wäre, daß sich der Beta-Koeffizient verringerte, obwohl die strukturelle Beziehung zwischen den beiden Variablen gleich geblieben wäre. Die Abhängigkeit des Beta-Koeffizienten von der in der jeweiligen Population oder Stichprobe erreichten Varianz ist also vor allem zu bedenken, wenn die Untersuchungsergebnisse verschiedener Studien für verschiedene Populationen oder Zeiträume untereinander verglichen werden.

Über den Aussagegehalt der nicht-standardisierten und der standardisierten Koeffizienten (zu den letztgenannten gehört auch Pearsons Korrelationskoeffizient r) ist viel debattiert worden (siehe z. B. Blalock 1971; Kim/Ferree 1981; King 1986; Kühnel 1985; Darlington 1968; Achen 1982). Dabei handelt es sich gelegentlich um überflüssige Gefechte, weil nicht beachtet wird, daß standardisierter und nicht-standardisierter Regressionskoeffizient den Begriff der »relativen Einflußstärke« bzw. der »theoretischen Bedeutung« jeweils anders auslegen, eher unterschiedliche als miteinander konkurrierende Aspekte behandeln. Der nicht-standardisierte Steigungskoeffizient antwortet auf die Frage: welcher durchschnittliche Niveauzuwachs ist in einer Variablen Y zu erwarten, wenn sich das Niveau einer Variablen X um eine Skaleneinheit erhöht? Der Beta-Koeffizient antwortet auf eine ebenfalls legitime, aber andere Frage, nämlich:

Wie kommt es, daß sich die einzelnen Untersuchungseinheiten hinsichtlich ihrer Y-Werte unterscheiden; wie läßt sich erklären, daß z. B. die SPD in den einzelnen Wahlkreisen so unterschiedliche Stimmenanteile bekommen hat? Die Antwort auf die zweite Frage sucht man in dem Ausmaß, in dem die Untersuchungseinheiten unterschiedliche Werte in den Regressorvariablen aufweisen. Die erste Frage setzt am Niveau, die zweite an der Streuung der Variablen an. Beide Aspekte sind nicht aufeinander reduzierbar. Vielleicht kann man sagen, daß die Antwort auf die erste Frage eine universellere, theoretisch allgemeinere Gültigkeit beanspruchen kann (falls das Modell korrekt spezifiziert wurde), während die Antwort auf die zweite Frage populations- bzw. stichprobenspezifisch ist.

Um die Gemüter weiter zu verwirren, ist auch noch ein **partieller Korrelationskoeffizient** konzipiert worden. Definiert ist er als reine Residuenkorrelation. Angenommen, das Modell sei (in Stichprobennotation):

$$(11\text{-}22) \qquad Y = a + b_1X_1 + b_2X_2 + \ldots + b_KX_K + e$$

Gesucht wird der partielle Korrelationskoeffizient $r_{yx(2).x(1)x(3)\ldots x(K)}$ für den Zusammenhang zwischen Y und X_2 unter »Konstanthalten« aller anderen Einflußfaktoren X_1, X_3, ..., X_K (die im Index nach dem Punkt aufgezählt werden). Man erhält diesen Koeffizieten, indem man zunächst die Residuen YR aus der Regression

$$(11\text{-}23) \qquad Y = a + b_1X_1 + b_3X_3 + \ldots + b_KX_K + e_Y$$

bildet: $YR = e_Y$. Sodann ermittelt man die Residuen $X_2R = e_{X(2)}$ aus

$$(11\text{-}24) \qquad X_2 = a + b_1X_1 + b_3X_3 + \ldots + b_KX_K + e_{X(2)}$$

Schließlich errechnet man den Pearson Produkt-Moment-Korrelationskoeffizienten für YR und X_2R wie in Teil I, Abschnitt 4.2.4 erläutert. Alternativ dazu ermittelt man den Determinationskoeffizienten R^2 aus der Regression

$$(11\text{-}25) \qquad YR = a + bX_2R + e$$

Die Wurzel aus diesem R^2 ist der gesuchte partielle Korrelationskoeffizient: $\sqrt{R^2} = r_{YX(2).X(1),X(3),\ldots,X(K)}$. Falls nur eine Kontrollvariable vorhanden ist, läßt sich der partielle Korrelationskoeffizient nach folgender Formel berechnen

$$(11\text{-}26) \qquad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

Der Einfachheit wegen haben wir die Kontrollvariable mit »2«, die zu kontrollierenden Variablen mit »1« und »3« indiziert.

Aus der Formel (11-26) geht hervor, daß der partielle Korrelationskoeffizient auch größer sein kann als derjenige der bivariaten Beziehung, z. B. dann, wenn abhängige und unabhängige Variable (hier: 1 und 3) positiv untereinander korrelieren, die Kontrollvariable (2) aber negativ mit der einen und positiv mit der anderen dieser beiden Variablen korreliert. Wenn die Kontrollvariable nicht konstant gehalten wird, wirkt sie in einem solchen Fall als »Suppressor« der Beziehung zwischen den beiden anderen Variablen (siehe Teil I, Abschn. 5.2.4).

Im Gegensatz zum Betakoeffizienten ist der partielle Korrelationskoeffizient eine symmetrische Maßzahl, d.h., man muß sich nicht darauf festlegen, welche der beiden Variablen als »abhängige« und welche als »unabhängige« gelten soll. Weitere Erläuterungen zum Unterschied zwischen partiellem Korrelationskoeffizient und Beta-Koeffizient gibt Holm (1977, S. 41-49). Hier sei nur noch darauf hingewiesen, daß im allgemeinen die Summe der quadrierten partiellen Korrelationskoeffizienten nicht identisch ist mit dem Determinationskoeffizienten, also dem quadrierten multiplen Korrelationskoeffizienten R^2. Die partiellen Korrelationskoeffizienten drücken nur die Varianzanteile aus, die die einzelnen Regressorenvariablen jeweils mit der abhängigen Variablen teilen. Varianzanteile, die von mehreren Regressorvariablen gemeinsam mit der Y-Variable geteilt werden, bleiben dabei (anders als beim Determinationskoeffizienten) außer acht.

11.4 Interaktive (multiplikative) Beziehungen

Bei der Darstellung der dreidimensionalen Tabellenanalyse in Teil I, Kap. 5 dieses Grundkurses haben wir bereits zwischen additiven und interaktiven Variablenbeziehungen unterschieden. Ein interaktives Beziehungsmuster war dadurch gekennzeichnet, daß in den Teiltabellen unterschiedlich starke Zusammenhänge zwischen einer (unabhängigen) Variablen X und einer (abhängigen) Variablen Y beobachtet wurden, je nachdem, welcher Wert oder welche Kategorie einer Kontrollvariablen Z gleichzeitig realisiert war: die **bedingten** Assoziationskoeffizienten in den einzelnen Partialtabellen waren so unterschiedlich, daß es nicht mehr sinnvoll

schien, sie auf dem Wege der Durchschnittsbildung zu einem gemeinsamen (nicht-bedingten) **partiellen** Koeffizienten zusammenzufassen. Ein Beispiel hierfür lieferte die Beziehung zwischen der Verfassungstradition (X) und der Links/Rechts-Orientierung (Y) der Abgeordneten der Frankfurter Nationalversammlung. Dieser Zusammenhang wurde durch die Konfessionszugehörigkeit (Z) der Abgeordneten spezifiziert: Die Verfassungstradition hatte bei Katholiken keinen, bei Protestanten aber einen relativ starken Einfluß auf das Abstimmungsverhalten, in dem sich die Links/Rechts-Orientierung manifestierte.

Bisher haben wir in der multiplen Regression nur additive Beziehungen betrachtet, Interaktionen waren in diesen Modellen nicht vorgesehen. Wir haben z. B. ermittelt, wie stark im Durchschnitt die SPD-Stimmenanteile ansteigen, wenn der Protestantenanteil (EV12) um 1 % zunimmt und der Grad der Industrialisierung und Urbanisierung konstant gehalten wird. Der entsprechende Steigungskoeffizient für EV12 war, wegen des »Konstanthaltens« der übrigen Regressorvariablen, zwar als **partieller** Koeffizient zu interpretieren, er galt jedoch »allgemein«; er war nicht spezifiziert für bestimmte Werte der anderen Regressorvariablen. Das additive Modell (11-14) sah z. B. nicht vor, daß er unterschiedliche Werte annehmen könnte, je nachdem welcher Urbanisierungsgrad erreicht wurde. Theoretisch (z. B. im Anschluß an die klassischen Arbeiten von Emile Durkheim) erscheint eine solche Hypothese jedoch durchaus als sinnvoll: unter städtischen Lebensbedingungen, der »Verdichtung« sozialer Kommunikation, der Konfrontation mit alternativen Deutungssystemen weichen konfessionelle Bindungen auf und determinieren in geringerem Maße als unter ländlichen Lebensbedingungen die politische Ausrichtung. Wenn diese These getestet werden soll, muß ein Regressionsmodell spezifiziert werden, das zuläßt, daß der Steigungskoeffizient der Konfessionsvariable (also das Gewicht ihres Einflusses auf das Votum für oder gegen die SPD) mit zunehmendem Urbanisierungsgrad geringer wird. Wie könnte eine solche Regressionsgleichung aussehen?

Um die Darstellung zu vereinfachen, berücksichtigen wir im folgenden nur einen der beiden Modernisierungsindikatoren, den Urbanisierungsgrad. Das additive Modell, in dem keine Interaktion vorgesehen ist, lautet (in Stichprobennotation):

$$(11\text{-}27) \qquad Y = a + b_1X_1 + b_2X_2 + e$$

mit Y = SPD-Stimmenanteil, X_1 = Protestantenanteil (EV12) und X_2 = Urbanisierungsgrad (URBAN). Die obige Interaktionshypothese können wir formal in folgender Gleichung ausdrücken:

(11-28) $b_1 = c + dX_2$

(Gemäß unserer Hypothese erwarten wir für "d" ein negatives Vorzeichen)

Durch diese Gleichung wird eine **lineare** Abhängigkeit des X_1 zugeordneten Steigungskoeffizienten von den Werten der zweiten Regressorvariable postuliert. Denkbar wären auch nicht-lineare Abhängigkeitsstrukturen, die wir hier aber nicht betrachten wollen. Gleichung (11-28) können wir in Gleichung (11-27) einsetzen:

$$(11\text{-}29) \quad Y = a + (c + dX_2)X_1 + b_2X_2 + e$$
$$= a^* + cX_1 + b_2^*X_2 + d(X_1X_2) + e$$

Neu an diesem Modell ist der multiplikative Ausdruck X_1X_2 in der letzten Zeile der Gleichung. Er wird dadurch gebildet, daß man für jeden einzelnen Fall (hier Wahlkreis) die jeweiligen Werte für EV12 und URBAN miteinander multipliziert (in SPSS^x mit Hilfe des COMPUTE-Statements, das man vor das REGRESSION-Kommando setzt). Damit wird eine neue Variable kreiert, die man zusätzlich in die Regressionsgleichung aufnimmt. Sämtliche Regressionskoeffizienten werden dann verfahrenstechnisch in gleicher Weise wie im rein additiven Modell nach der üblichen Kleinstquadratemethode (OLS) geschätzt. Da mit X_1X_2 eine neue Variable hinzugekommen ist, unterscheiden sich im allgemeinen die Regressionskoeffizienten des multiplikativen von den entsprechenden Koeffizienten des additiven Modells; deshalb sind sie in (11-29) mit einem Sternchen versehen worden. Der Einfachheit wegen verwenden wir aber für das multiplikative Modell im folgenden wieder die vertrauten Symbole (in Stichprobennotation):

$$(11\text{-}30) \quad Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2 + e$$

behalten aber »im Kopf«, daß a, b_1 und b_2 in (11-30) nicht identisch sind mit den ebenso bezeichneten Koeffezienten in (11-27). Gleichung (11-30) ist die übliche Form, in der man interaktive Beziehungen in einem Regressionsmodell darstellt[8]. Der multiplikative Ausdruck sorgt dafür, daß die Steigungskoeffizienten für X_1 **und** für X_2 von der jeweils anderen Regres-

[8] Zu verschiedenen Möglichkeiten, diese oder andere Gleichungsformen mit theoretisch unterschiedlichen Interaktionsmodellen zu verknüpfen, siehe Southwood 1978.

sorvariable abhängig sind. Dies können wir leicht überprüfen, indem wir zu Gleichung (11-30) die partiellen ersten Ableitungen bilden (auch im additiven Modell erhalten wir formal die Steigungskoeffizienten mit Hilfe partieller Ableitungen):

$$(11\text{-}31) \qquad \partial Y/\partial X_1 = b_1 + b_3 X_2$$

$$(11\text{-}32) \qquad \partial Y/\partial X_2 = b_2 + b_3 X_1$$

Der Koeffizient b_3 gibt also an,

a) um welchen Betrag sich der Steigungskoeffizient für X_1 ändert, wenn X_2 um eine Einheit zunimmt (für unser Beispiel vermuten wir für b_3 ein negatives Vorzeichen, da wir annehmen, daß das Regressionsgewicht für die Konfessionsvariable mit steigendem Urbanisierungsgrad abnimmt),
b) um welchen Betrag sich der Steigungskoeffizient für X_2 ändert, wenn X_1 um eine Einheit zunimmt.

Der multiplikative Term in der Regressionsgleichung formalisiert die Hypothese der Interaktion zweier »unabhängiger« Variablen in ihrer Wirkung auf eine »abhängige« Variable somit als »symmetrische« Beziehung, wie wir das ja auch in der Tabellenanalyse kennen gelernt haben. Wenn die Konfessionsvariable mit zunehmendem Urbanisierungsgrad einen (betragsmäßig) kleineren Steigungskoeffizienten erhält, impliziert das auch für den Urbanisierungsgrad einen Steigungskoeffizienten, der in gleicher Weise mit dem zunehmenden Protestantenanteil abnimmt - auch wenn man diese »Seite« der Interaktionsbeziehung nicht kausal deuten möchte. Diese Implikation wird durch folgende Überlegungen verständlich:

(a) Aus der Feststellung, auf der »Aggregatebene« der Wahlkreise bestehe eine positive Beziehung zwischen Protestantenanteil und SPD-Stimmenanteil, folgt nicht notwendig, daß auch auf der »Individualebene« die protestantischen Wahlberechtigten mit einer größeren Wahrscheinlichkeit für die SPD votieren als die katholischen[9]. Gestützt auf andere historische Kenntnisse wollen wir dies jedoch annehmen.
(b) Das Schätzergebnis für das rein additive Regressionsmodell können wir demnach wie folgt interpretieren: Bei zunehmendem Urbanisierungsgrad steigt sowohl für Katholiken als auch für Protestanten die

---

[9] Diese Problematik ist in der Literatur ausführlich unter dem (mißverständlichen) Stichwort des »ökologischen Fehlschlusses« behandelt worden. Zur Einführung siehe z. B. Hummell (1972).

Wahrscheinlichkeit eines SPD-Votums, wobei (wegen der Additivität der Beziehung) die Wahrscheinlichkeitsdifferenz zwischen den beiden Personengruppen gleich bleibt.

(c) Die in Gleichung (11-28) für die Aggregatebene formulierte Interaktionshypothese mit einem negativen Koeffizienten für den multiplikativen Term ist demgegenüber aus der Vermutung abgleitet, daß sich auf der Individualebene die Wahrscheinlichkeiten für ein SPD-Votum bei Katholiken und Protestanten einander **annähern**, wenn der Urbanisierungsgrad zunimmt. Je stärker sich die beiden Individualwahrscheinlichkeiten annähern, um so geringer ist nämlich auf der Aggregatebene (betragsmäßig) der Effekt (d.h. der Steigungskoeffizient) des Protestantenanteils (oder, bei umgekehrter Kodierung, des Katholikenanteils). Dies gilt auch umgekehrt: Je geringer (betragsmäßig) die Steigung der Aggregatregression, desto geringer die entsprechende Wahrscheinlichkeitsdifferenz auf der Individualebene.

(d) Bei der gegebenen Ausgangslage (Urbanisierungsgrad wirkt positiv auf SPD-Stimmenanteil; Protestanten wählen eher als Katholiken die SPD) kann die Annäherung der Individualwahrscheinlichkeiten bei zunehmender Urbanisierung nur heißen: Die Wahrscheinschlichkeit, SPD zu wählen, muß sich unter dem Einfluß zunehmender Urbanisierung bei den Katholiken stärker erhöhen als bei den Protestanten.

(e) Auf der Aggregatebene bedeutet dies: Je höher der Anteil der Katholiken, um so stärker der Effekt der Urbanisierung zugunsten des SPD-Stimmenanteils - oder, anders formuliert: Je größer der Anteil der Protestanten, um so geringer der Urbanisierungseffekt. Der formalen Symmetrie des multiplikativen Interaktionstherms muß also nicht unbedingt eine kausale Symmetrie entsprechen. (Theoretisch ist es wohl wenig sinnvoll anzunehmen, eine Zunahme des Protestantenanteils »bewirke« eine Schwächung des Urbanisierungseffekts.)

Auch die Interpretation der Koeffizienten b_1 und b_2 in Gleichung (11-30) ergibt sich aus den Gleichungen (11-31) und (11-32). Der Koeffizient b_1 gibt an, um welchen Betrag sich Y ändert, wenn X_1 um eine Einheit zunimmt **und** gleichzeitig $X_2 = 0$ ist. Entsprechend gibt b_2 den (bedingten) Steigungskoeffizienten für X_2 unter der Bedingung an, daß $X_1 = 0$ ist. Diese Bedingungen, $X_1 = 0$ und $X_2 = 0$, können, wie in unserem Beispiel, außerhalb des beobachteten Wertebereichs liegen. Die unter diesen Bedingungen »extrapolierten« Steigungskoeffizienten mögen deshalb, wie der Ordinatenabschnitt a, »unrealistische« Schätzgrößen darstellen. Das mindert aber nicht zwangsläufig die Validität der Steigungskoeffizienten, die unter anderen X_1- bzw. X_2-Bedingungen mit Hilfe von b_1 bzw. b_2 ermittelt wer-

den (siehe die Beispielrechnungen unten). Wichtig ist, überhaupt zu verstehen, daß es sich bei den Steigungskoeffizienten $\delta y/\delta x_1$ bzw. $\delta y/\delta x_2$ im multiplikativen Modell immer nur um **bedingte** Koeffizienten handelt, die mit den Werten der jeweils anderen Regressorvariable variieren. Falls eine interaktive Beziehung vorliegt, ist es also unsinnig, inhaltlich zwischen additiven und interaktiven, den sog. »Haupt«- und »Nebeneffekten« zu trennen. Das Regressionsgewicht, der Einfluß einer bestimmten Variablen X_1 ist (falls Interaktion vorliegt) eben nicht allgemein gültig durch eine Konstante anzugeben, sondern nur spezifisch unter der Bedingung eines bestimmten Wertes von X_2[10].

Nachdem wir dieses formale Modell interaktiver Beziehungen (Gleichung (11-30)) erklärt haben, wollen wir nun dessen Parameter schätzen. Zur Wiederholung: X_1 = EV12 (Protestantenanteil), X_2 = URBAN (Urbanisierungsgrad) und X_1X_2 = URBEV12 (Produkt der beiden unabhängigen Variablen). Abb. 11.6 stellt die Ergebnisse sowohl des additiven Modells (11-27) wie auch des multiplikativen Modells (11-30) zusammen.

Zunächst einmal läßt sich feststellen, daß der multiplikative Ausdruck hoch signifikant ist. Das zeigt sowohl der t-Test (mit t = 7.99) für das Regressionsgewicht b_3 = .004002 als auch der F-Test mit »F Change« = 63.855 für den Zuwachs erklärter Varianz in Höhe von »R Square Change« = .04066. Das Vorzeichen von b_3 ist jedoch nicht, wie erwartet, negativ, sondern positiv. Das bedeutet, daß entgegen unserer Hypothese das Gewicht der Konfessionszugeörigkeit für die parteipolitische Ausrichtung der Wähler mit wachsender Urbanisierung nicht ab-, sondern zunimmt. Ad hoc könnte man dies damit »erklären«, daß unter den Bedingungen städtischen Lebens die Bedrohung der eigenen kulturellen Identität durch Angebot und Anspruch alternativer Deutungssysteme und Lebensweisen zunimmt und daß dies zu einer defensiven Reaktion führt, in der sich die Herkunftsmilieus zunächst stärker voneinander abschotten. Aber diese Alternativ-Hypothese ist post festum formuliert und bedarf anderer Forschungsdaten, um gestützt zu werden.

Dem Ergebnisausdruck ist zu entnehmen, daß die b-Koeffizienten für EV12 (b_1 = .06) und URBAN (b_2 = .15) im multiplikativen Modell erheblich niedriger sind als die entsprechenden Koffizienten (.28 und .39) im additiven Modell. Wir wissen aber aus den Gleichungen (11-31) und (11-32), daß es sich bei den erstgenannten um Steigungskoeffizienten handelt, die für die unrealistische Bedingung extrapoliert sind, daß der Protestantenanteil bzw. der Urbanisierungsgrad bei Null liegt. Um das sehr

[10] In Modellen mit mehr als zwei Regressorvariablen kann eine Variable mit mehreren anderen Variablen interagieren; außerdem können mehr als zwei Variablen gleichzeitig interargieren (Interaktion »höherer Ordnung«); siehe hierzu Friedrich (1982, S. 829 ff.).

niedrige Regressionsgewicht ($b_3 = .004$) für den multiplikativen Ausdruck zu verstehen, muß man sich klar machen, daß durch die Multiplikation der beiden Prozentanteile die Skala auf Tausendereinheiten ausgedehnt wurde, der Y-Zuwachs pro X_1X_2 - Einheit also entsprechend niedriger sein muß.

In welchen Grenzen variieren nun die Steigungskoeffizienten für die Konfessionsvariable unter den Bedingungen unterschiedlicher Urbanisierungsgrade? Der niedrigste Urbanisierungsgrad in den Wahlkreisen von 1912 liegt bei $X_2 = 7$ %, der höchste bei 100 %. Demgemäß schwanken die »realistischen« Steigungskoeffizienten für EV12 gemäß Gleichung (11-31) zwischen

$$(11\text{-}33) \qquad b_{min}(EV12) = b_1 + b_3X_2$$
$$= 0{,}06 + 0{,}004 \cdot 7 = 0{,}088$$

und

$$(11\text{-}34) \qquad b_{max}(EV12) = 0{,}06 + 0{,}004 \cdot 100 = 0{,}46$$

Der Urbanisierungsgrad hat im arithmetischen Mittel einen Wert von 61 %. Unter dieser Bedingung ist der Steigungskoeffizient für die Konfessionsvariable

$$(11\text{-}35) \qquad b_x(EV12) = 0{,}06 + 0{,}004 \cdot 61 = 0{,}30$$

Er weist an dieser Stelle etwa die gleiche Größe auf wie der generelle Steigungskoeffizient $b_1 = .28$ im additiven Modell (11-27). In gleicher Weise ließe sich auch die Bandbreite der Steigungskoeffizienten für die Urbanisierungsvariable unter wechselnden Werten der Konfessionsvariable errechnen. Wir wollen uns dies aber ersparen und lediglich noch eine Prognoserechnung ausführen: den Erwartungswert des SPD-Stimmenanteils unter der Bedingung ausrechnen, daß $X_1 = \bar{x}_1 = 63$ und $X_2 = \bar{x}_2 = 61$:

(11-36)

$$\hat{y} = a + b_1(EV12) + b_2(URBAN) + b_3(URBEV12)$$
$$= 0{,}714 + 0{,}06 \cdot 63 + 0{,}15 \cdot 61 + 0{,}004 \cdot 63 \cdot 61$$
$$= 29{,}016$$

Dieser Wert deckt sich bis auf Rundungsfehler mit dem beobachteten arithmetischen Mittel der abhängigen Variablen. (Wir hatten bereits in

Kapitel 10 darauf hingewiesen, daß die Regressionskurve bzw. Regressionsfläche durch den Schnittpunkt der Mittelwert-Koordinaten aller Variablen verläuft.) In gleicher Weise können Erwartungswerte für Y auch unter anderen Bedingungen ausgerechnet und miteinander verglichen werden.

Zum Schluß sei noch darauf hingewiesen, daß neben metrischen auch dummy-kodierte qualitative Variablen als Prädiktoren in ein Regressionsmodell aufgenommen werden können. Solche Modelle bezeichnet man als »Kovarianzmodelle«. Indem multiplikative Terme aus einer metrischen Variablen und einer dummy-kodierten Variablen gebildet werden, läßt sich prüfen, ob der Steigungskoeffizient der metrischen Prädiktorvariable in den verschiedenen Gruppen von Untersuchungseinheiten, die durch die Kategorien der qualitativen Variablen definiert sind, unterschiedlich ist oder nicht. Falls keine multiplikativen Terme gebildet werden, wird lediglich geprüft, ob sich die Y-Mittelwerte in den verschiedenen Gruppen »signifikant« voneinander unterscheiden (siehe Pyndick/Rubinfeld 1981, S. 111-114, 135-137).

11.4.1 Zur Diskussion über die Anwendbarkeit multiplikativer Modelle (*)

Multiplikative Regressionsmodelle sind von Sozialwissenschaftlern häufig deshalb verworfen worden, weil zwischen den Ursprungsvariablen und dem Produktterm in der Regel eine hohe Multikollinearität besteht (wie auch in unserem Beispiel). Wie Friedrich (1982) gezeigt hat, läßt dieser Typ der Multikollinearät jedoch keine fatalen Konsequenzen für die Güte der Parameterschätzungen und der Signifikanztests erwarten (vgl. Miller/Farmer 1988). Das hängt damit zusammen, daß auch die Standardfehler der Parameter eines multiplikativen Modells bedingte Größen sind. Die im Ergebnisausdruck notierten Standardfehler für b_1 und b_2 beschreiben im interaktiven Modell die (geschätzte) Stichprobenstreuung der Steigungskoeffizienten nur unter der Bedingung, daß X_2 respektive X_1 den Wert Null annehmen. Unter anderen Bedingungen müssen, wie wir sahen, die Koeffizienten b_1 und b_3 bzw. b_2 und b_3 miteinander kombiniert werden, um den bedingten Steigungskoeffizienten errechnen zu können. Unter Anwendung der Regeln für das Rechnen mit Kovarianzen (siehe Teil I, Abschnitt 4.2.6) lassen sich (unter Annahme korrekter Modellspezifikation) die Standardfehler für die **bedingten** Steigungskoeffizienten allgemein nach folgenden Formeln berechnen (siehe Friedrich 1982, S. 810):

(11-37) $$\hat{\sigma}_{b(1)+b(3)X(2)} = \sqrt{\text{var}(b_1)+X_2^2\text{var}(b_3)+2X_2\text{cov}(b_1,b_3)}$$

$$\hat{\sigma}_{b(2)+b(3)X(1)} = \sqrt{\text{var}(b_2)+X_1^2\text{var}(b_3)+2X_1\text{cov}(b_2,b_3)}$$

Die Varianzen und Kovarianzen der einzelnen b-Koeffizienten sind dem Ergebnisausdruck (s. Abb. 11.6) zu entnehmen. Sie stehen in der Diagonalen bzw. dem unteren Dreieck der »Var-Covar Matrix of Regression Coefficients«. Zur Illustration rechnen wir den Standardfehler des Steigungskoeffizienten für EV12 unter der Bedingung aus, daß der Urbanisierungsgrad 80 % beträgt:

(11-38)
$$\begin{aligned}\hat{\sigma}_{b(1)+b(3)X(2)} &= \sqrt{0{,}0009237 + 80 \cdot 0{,}0000002508} \\ &\quad + 2 \cdot 80 \cdot (-{,}00001362) \\ &= 0{,}0187\end{aligned}$$

Dieser bedingte Standardfehler des bedingten Steigungskoeffizienten für EV12 ist kleiner als der im Ergebnisausdruck wiedergegebene Standardfehler $s_b = 0.030392$, der für die Bedingung gilt, daß URBAN = 0 ist. Er ist auch nur geringfügig größer als $s_b = 0.014632$ bezüglich EV12 im additiven Modell. Da die Schätzer b_1 und b_3 bzw. b_2 und b_3 im multiplikativen Modell häufig stark negativ miteinander korrelieren (bei positiver Korrelation zwischen den Ursprungsvariablen und dem Produktterm), ist der Standardfehler für die **bedingten** Steigungskoeffizienten bei bestimmten Werten von X_2 bzw. X_1 sogar niedriger als der entsprechende Standardfehler im additiven Modell.

Die im Ergebnisausdruck notierten t-Werte für X_1 und X_2 (hier 1.989 und 4.474) können bei der Schätzung multiplikativer Modelle häufig einen Wert $t < 2$ annehmen. Dies allein ist noch kein Grund, das Modell abzulehnen bzw. durch Eliminieren der vermeintlich nicht-signifikanten Komponenten zu modifizieren. Denn dieser t-Wert gibt das Verhältnis zwischen Steigungskoeffizient und Standardfehler nur für den Fall an, daß die andere Regressorvariable den Wert Null aufweist. Man sollte aber den Steigungskoeffizienten und seinen Standardfehler auch unter anderen Bedingungen ermitteln und ins Verhältnis zueinander setzen. Wenn auch die Bedingungen, die den Standardfehler minimieren, keine signifikanten t-Werte ergeben, ist allerdings eine Modellrevision geboten.

Die X_1- und X_2-Werte, bei denen die Standardfehler minimal sind, erhält man, indem man die Gleichungen für die bedingten Standardfehler (siehe 11-37)) differenziert, die ersten Ableitungen gleich Null setzt und nach X_1 und X_2 auflöst (siehe Friedrich 1982, S. 812):

(11-39) $X_2 = -\text{cov}(b_1,b_2)/\text{var}(b_3)$ für den Steigungskoeff. von X_1

(11-40) $X_1 = -\text{cov}(b_2,b_3)/\text{var}(b_3)$ für den Steigungskoeff. von X_2

In unserem Beispiel ergibt sich der minimale Standardfehler für den Steigungskoeffizienten von EV12 (X_1) unter der Bedingung

(11-41)

$$\text{URBAN}(X_2) = -(-0,00001362)/0,0000002508) = 54,3$$

Unter dieser Bedingung nimmt der Steigungskoeffizient für EV12 folgenden Wert an (siehe Gleichung (11-31):

(11-42)
$$\begin{aligned} b(\text{EV12}\,|\,\text{URBAN}=54,3) &= b_1 + b_3(54,3) \\ &= 0,06 + 0,004 \cdot 54,3 \\ &= 0,2772 \end{aligned}$$

Sein Standardfehler (siehe Gleichung (11-37)) beträgt

(11-43)
$$\begin{aligned} \hat{\sigma}_{b(1)} &= \sqrt{\text{var}(b_1) + X_2^2\text{var}(b_3) + 2X_2\text{cov}(b_1,b_3)} \\ &= \sqrt{0,0009237 + (54,3)^2(0,0000002508) \\ &\quad + 2 \cdot 54,3(-0,00001362)} \\ &= \sqrt{0,0001848} = 0,01359 \end{aligned}$$

Dieser Standardfehler ist noch niedriger als der Standardfehler des generellen Steigungskoeffizienten von EV12 im additiven Modell s_b = .014632. Dies gilt allgemein: der minimale Standardfehler für einen bedingten Steigungskoeffizienten im multiplikativen Modell ist (trotz Multikollinearität) stets geringer als der Standardfehler des allgemeinen Steigungskoeffizienten im additiven Modell (siehe Friedrich 1982, S. 813 f.).

Unter der Bedingung URBAN=54.3 ergibt sich aus Gleichungen (11-42) und (11-43) somit folgender t-Wert für den bedingten Steigungskoeffizienten von EV12:

$$(11\text{-}44) \qquad t = \frac{b_1}{\hat{\sigma}} = \frac{0{,}2772}{0{,}01359} = 20{,}4$$

Er ist also erheblich höher als der t-Wert (1.989), der im Ergebnisausdruck des multiplikativen Modells für b_1 ausgewiesen ist und nur unter der Bedingung gilt, daß URBAN = 0 ist. »Not until conditional slopes and t tests are calculated within the observed range of experience of the variables can valid conclusions be drawn. Statistically insignificant b_1's, b_2's, and b_3's may nevertheless combine to produce statistically significant conditional effects.« (Friedrich 1882, S. 821).

Ein weiterer, häufig vorgebrachter Einwand gegen die Spezifikation multiplikativer Modelle in den Sozialwissenschaften besagt, daß die (geschätzten) Modellparameter nur für Ratio-, nicht aber für Intervallskalen interpretierbar seien (siehe z. B. Althauser 1971; Allison 1977). Friedrich (1982, S. 821 ff.) zeigt einleuchtend, daß die diesbezüglichen Argumente nicht stichhaltig sind, wenn man den bedingten Charakter der Parameterschätzungen ernst nimmt.

Da, wie angedeutet, auch Multikollinearität in der Regel die Parameterschätzungen in multiplikativen Modellen nicht zu stark beeinträchtigt, sollte sich der Leser ermutigt fühlen, solche Modelle zu spezifizieren und zu schätzen. Gegen quantifizierende Analysen in der historischen Sozialwissenschaft haben Kritiker häufig geltend gemacht, man könne den Einfluß einer bestimmten Variable nicht quantifizieren, weil er »kontextabhängig« sei. Mit Kontextabhängigkeit ist aber meist nichts anderes gemeint als die Interaktion mehrerer Variablen hinsichtlich ihrer Wirkung auf eine andere Variable. Solche »Wechselbeziehungen« lassen sich jedoch, wie wir gezeigt haben, im Prinzip modellieren. Es bleibt natürlich die Frage, ob man den relevanten »Kontext« hinreichend gut durch entsprechende Daten erfassen kann.

11.5 Ausblick auf Pfadmodelle (*)

Das Wahlergebnis der SPD bei den Reichstagswahlen von 1912 hing in den verschiedenen Wahlkreisen sicherlich auch davon ab, wie stark die Arbeiterbewegung vor Ort organisiert war. Als Indikator hierfür läßt sich die Mitgliederzahl der Freien Gewerkschaften im Verhältnis zur Bevölkerungsgröße einsetzen. Je höher dieser Anteil, so können wir vermuten, desto stärker (in der Regel) die Unterstützung des SPD-Kandidaten im Wahlkampf und desto höher dessen Stimmenanteil bei der Wahl.

Das Streudiagramm in Abb. 11.7 zeigt, daß der Zusammenhang zwischen gewerkschaftlichem Oragnisationsgrad (GEWORG) und SPD-Stimmenanteil nicht linear ist. Eine »gedachte« Regressionslinie steigt zunächst, bei niedrigen GEWORG-Werten, steil an, flacht aber zunehmend ab; die Steigung tendiert dann gegen Null. Mit anderen Worten: Nimmt der Mitgliederanteil beispielsweise von 1 auf 2 Prozent der Bevölkerung zu, so hat dies einen wesentlich größeren positiven Effekt auf den SPD-Stimmenanteil als eine Zunahme des Mitgliederanteils von 10 auf 11 Prozent. Nicht-lineare Beziehungen wie diese können in eine lineare Beziehung transformiert werden, indem man die unabhängige Variable logarithmiert. (Näher erläutert werden derartige Transformationen in Kap. 12.1.) Statt GEWORG setzen wir den Logarithmus von GEWORG in die Regressionsgleichung ein:

$$(11\text{-}45)\quad PROZSPD = a + b_1EV12 + b_2URBAN + b_3URBEV12 + b_4(\log_{10}GEWORG) + e$$

In SPSS^x läßt sich eine Variable sehr leicht mit Hilfe eines COMPUTE-Statements und einer in das System eingebauten arithmetischen Funktion logarithmieren - entweder zur Basis »zehn« oder (als sog. »natürlicher« Logarithmus) zur Basis der Eulerschen Zahl e $\approx$ 2.718. Der größeren Anschaulichkeit wegen wählen wir hier den Zehnerlogarithmus. Die so transformierte Variable erhält den Namen GEWLOG:

```
COMPUTE GEWLOG=LG10(GEWORG)
```

Die Wahl der Log-Basis ist im Prinzip beliebig, da die entsprechenden Logarithmen zueinander proportional sind. Zu beachten ist allerdings, daß ein Logarithmus für Zahlen kleiner 0 nicht definiert ist. Um dieses Problem zu umgehen, wird gelegentlich vorgeschlagen, vor dem Logarithmieren allen Skalenwerten einen konstanten positiven Betrag hinzuzufügen. Im Falle von Ratioskalen ist dies jedoch nicht zulässig. In unserem Datensatz haben 88 Wahlkreise keine (meßbare) gewerkschaftliche Oganisation. Wir schließen sie im folgenden aus der Analyse aus; die Allgemeinheit unserer Untersuchungsergebnisse ist insoweit eingeschränkt. Die Gewichtung durch die Zahl der Wahlberechtigten wird entsprechend korrigiert und auf N=397 - 88=309 Fälle normiert (siehe Kap. 10.1). Nach dem Logarithmieren zeigt das Streudiagramm der Variablen PROZSPD und GEWLOG eine Punktewolke, die mit der Annahme einer linearen Beziehung verträglich ist.

Die Regressionskoeffizienten des Modells (11-45) werden mit dem gleichen Verfahren (OLS, Kleinstquadrateverfahren) geschätzt, das wir auch

bisher stets angewandt haben. Die Ergebnisse sind Abb. 11.8 zu entnehmen.

Die Interpretation der Steigungskoeffizienten für die miteinander interagierenden Variablen EV12 und URBAN ist in Abschnitt 11.4 ausführlich erläutert worden. Neu ist die Interpretation des Steigungskoeffizienten einer logarithmierten Variable. GEWLOG ist eine additive Modellkomponente; der Regressionskoeffizient $b_4 = 7.6$ gibt also an, um wieviel (Prozent-)Einheiten der SPD-Stimmenanteil durchschnittlich wächst, wenn GEWLOG um eine Einheit zunimmt (und die anderen Regressorvariablen konstant gehalten werden). Wächst der Zehnerlogarithmus um eine Einheit, impliziert dies in der Originalvariablen (GEWORG) eine Zunahme um den Faktor 10. Also bedeutet $b_4 = 7.6$: wenn sich der Anteil der Gewerkschaftsmitglieder verzehnfacht, steigt der SPD-Stimmenanteil um 7.6 % an. Dies läßt sich anhand einfacher Beispielrechnungen veranschaulichen:

Der gewerkschaftliche Mitgliederanteil (im folgenden: X_4) schwankt in den verschiedenen Wahlkreisen von 0 bis knapp 16 % der Bevölkerung. Wenn er von 0.5 % auf $10 \cdot 0.5\,\% = 5\,\%$ steigt, können wir, laut obigem Modell, einen SPD-Stimmenzuwachs von 7.6 Prozentpunkten erwarten. Der gleiche Zuwachs ist auch dann zu erwarten, wenn X_4 von 1 auf 10 % oder von 1.5 auf 15 % ansteigt. Natürlich läßt sich der erwartete Zuwachs an SPD-Stimmenanteilen (im folgenden: Y) auch für jede andere Veränderung des gewerkschaftlichen Organisationsgrades (X_4) ermitteln, indem man Gleichung (11-45) zur Prognose benutzt und die entsprechenden bedingten Erwartungswerte für Y ausrechnet. Eine Zunahme in X_4 von 2 % auf 3 % z. B. bewirkt in der Zehnerlog-Skala GEWLOG einen Zuwachs von $\log 2 = 0.3013$ auf $\log 3 = 0.4771$. Da GEWLOG nicht mit EV12 (X_1) und URBAN (X_2) interagiert, können wir diese beiden Variablen der Einfachheit wegen in (11-45) gleich Null setzen und erhalten somit

(11-46)

$$E(Y \mid \text{GEWLOG}=0{,}3013) = 8{,}969 + 7{,}6 \cdot 0{,}3013 = 11{,}259$$
$$E(Y \mid \text{GEWLOG}=0{,}4771) = 8{,}969 + 7{,}6 \cdot 0{,}4771 = 12{,}599$$

Die Erwartungswerte der SPD-Stimmenanteile nehmen also um 12.599 - 11.259 = 1.34 Prozent zu, wenn sich der Anteil der Gewerkschaftsmitglieder von zwei auf drei Prozent der Bevölkerung erhöht. Die gleiche Zuwachsgröße erhält man durch folgende Rechnung: $(\log 3 - \log 2) \cdot b_4 = 0{,}1758 \cdot 7{,}6 = 1{,}336$.

Ein Nachteil des Regressionsmodells (11-45) besteht darin, daß es alle unabhängigen Variablen hinsichtlich ihres Einflusses auf die abhängige Variable kausaltheoretisch auf die gleiche Stufe stellt, ihnen lediglich im

Schätzvorgang unterschiedliche Gewichte (Steigungskoeffizienten) zuweist. Tatsächlich dürfte jedoch GEWORG bzw. GEWLOG gegenüber den beiden anderen Regressorvariablen EV12 und URBAN kausal nachgeordnet sein. Wir können annehmen, daß Protestantenanteil und Urbanisierungsgrad positiv auf den gewerkschaftlichen Organisationsgrad wirken. Eine solche Kausalhypothese läßt sich graphisch in Form eines sog. **Pfaddiagramms** (siehe Abb. 11.9) darstellen (wir eliminieren zunächst den multiplikativen Term, fügen ihn aber später wieder hinzu).

Jede Variable, die eine andere Variable beeinflußt, steht links von dieser. Die Richtung des kausalen Einflusses ist durch Pfeile symbolisiert, die von der unabhängigen auf die abhängige Variable zielen. Den Pfeilen sind Koeffizienten (»Pfadkoeffizienten«) beigeordnet, die die »Stärke« der Beziehung (des »kausalen Effekts«) ausdrücken sollen; sie sind zu schätzen. In der Terminologie der Pfadanalyse bezeichnet man die ganz links stehenden Variablen (hier EV12 und URBAN) als **exogene** Variablen. Von ihnen nimmt man an, daß sie durch keine andere Variable des analysierten Systems beeinflußt sind. Sie sind mit gerundeten Doppelpfeilen verbunden, die darauf hinweisen, daß zwischen ihnen eine Korrelation besteht (bestehen kann), die aber nicht kausal gedeutet wird. GEWLOG ist in unserem Pfadmodell eine **intervenierenden** Variable, weil sie sowohl als abhängige Variable (im Verhältnis zu EV12 und URBAN) wie auch als unabhängige Variable (im Verhältnis zu PROZSPD) auftritt. Intervenierende und exogene Variablen bezeichnet man in bezug auf kausal nachgeordnete Variablen als **prädeterminiert** (siehe Opp/Schmidt 1976, S.98); abhängige Variablen bezeichnet man in bezug auf die prädeterminierten Variablen als **endogen**. Intervenierende Variablen sind sowohl endogen (in bezug auf die kausal vorgeordneten Variablen) als auch prädeterminiert (in bezug auf Variablen, die ihnen in der Kausalkette noch folgen). Pfadmodelle, in denen sämtliche (einfache) Pfeile nur von links nach rechts verlaufen, in denen also keine »Feedback«-Beziehungen (Wechselwirkungen von X_k nach X_m und von X_m nach X_k) auftreten, nennt man **rekursive** Kausalsysteme[11]. Wenn Feedback-Beziehungen vorliegen, spricht man von nicht-rekursiven Kausalsystemen, die wir hier aber nicht betrachten wollen (siehe Asher 1983, S. 53ff.), obwohl eine solche Beziehung in unserem Beispiel zwischen Wahlerfolg der SPD und gewerkschaftlichem Organisationsgrad nicht auszuschließen ist.

Ein Pfadmodell läßt sich in Form eines Gleichungssystems, einer Menge sog. **Strukturgleichungen**, darstellen, wobei für jede abhängige Variable eine Gleichung angegeben wird. Das Pfadmodell in Abb. 11.9 wird durch zwei Strukturgleichungen formuliert (in Stichprobennotation):

[11] Einige Autoren (z.B. Bollen 1989, S.81) führen die Unkorreliertheit der Fehlervariablen zwischen den einzelnen Gleichungen (siehe unten) als zusätzliches Definitionsmerkmal ein.

$$(11\text{-}47\ a) \qquad Y = a_1 + b_1X_1 + b_2X_2 + b_4(\log X_4) + e_1$$

$$(11\text{-}47\ b) \qquad \log X_4 = a_2 + b_5X_1 + b_6X_2 + e_2$$

Wenn man mit z-standardisierten Variablen, also mit Beta-Koeffizienten rechnen möchte (siehe Abschn. 11.3), wird daraus

$$(11\text{-}48\ a) \qquad Z_y = \beta_1Z_1 + \beta_2Z_2 + \beta_4Z_4 + u_1$$

$$(11\text{-}48\ b) \qquad Z_4 = \beta_5Z_1 + \beta_6Z_2 + u_2$$

Da durch diese Transformation die arithmetischen Mittel aller Variablen auf Null gesetzt sind, entfällt der Ordinatenabschnitt. Bei Z_4 handelt es sich um die z-standardisierte GEWLOG-Variable, die in (11-47 a u. b) als $\log X_4$ bezeichnet ist. Hinsichtlich der anderen Variablen dürfte sich eine Erklärung zur Notation erübrigen. Den Index »3« haben wir ausgelassen; er bleibt für den multiplikativen Term URBEV12 reserviert, den wir später einfügen werden. Üblicherweise (siehe Asher 1983, S. 31) schätzt man die Koeffizienten jeder dieser Gleichungen nach dem Kleinstquadrateverfahren (OLS-Regression)[12]. Zusätzlich zu den bisher bei der Regressionsanaylse schon gemachten Voraussetzungen wird angenommen, daß die Residuen e_1 und e_2 unkorreliert sind. Die Koeffizienten, die den einzelnen Pfaden, damit also den verschiedenen prädeterminierten Variablen zugeordnet sind, bezeichnet man als **direkte kausale Effekte**. Die schräg von außen auf die abhängigen Variablen geführten Pfeile symbolisieren den im Modell nicht erklärten »Einfluß« der Residualvariablen. Als Pfadkoeffizient ist hierfür die Wurzel des Unbestimmtheitsmaßes ($1-R^2$) eingesetzt.

---

[12] Siehe Asher (1983, S. 31). Falls das rekursive System nicht »vollständig« ist, d. h., wenn angenommen wird, daß in der diagrammatischen Übersetzung des Modells irgendwelche Pfeile zwischen einer der prädeterminierten und einer der abhängigen Variablen auszulassen sind, entstehen beim Schätzen und Testen gewisse Probleme, die wir in diesem kurzen Ausblick aber nicht behandeln wollen (siehe Asher 1983, S. 47 f.). Wir stützen uns hier auf den Hinweis von Bollen (1989, S. 115), daß allgemein in rekursiven Modellen (mit unkorrelierten Fehlern) OLS-Schätzer identisch sind mit sog. »Full-Information-Maximum-Likelihood« - Schätzern (siehe unten Kap. 12.2).

Wie in der Pfadanalyse üblich, tragen wir in das Pfaddiagramm zunächst die in Abschnitt 11.3 erläuterten **standardisierten** Regressionskoeffizienten, also die Beta-Koeffizienten ein (siehe Abb. 11.10)

Der Vorteil der Pfadanalyse gegenüber der »einfachen« Regressionsanalyse besteht vor allem darin, daß man nicht nur die »direkten«, sondern auch die »indirekten« und damit die »totalen« Effekte einer prädeterminierten Variable auf eine abhängige Variable berechnen kann. In unserem Pfadmodell haben wir 5 direkte Effekte notiert: Von EV12, URBAN und GEWLOG auf PROZSPD sowie von EV12 und URBAN auf GEWLOG. Sie sind jeweils mit »normaler« OLS-Regression als Parameter der Gleichungen (11-48 a) und (11-48 b) geschätzt worden. Unter inhaltlichen Aspekten ist hier festzuhalten, daß der gewerkschaftliche Organisationsgrad unabhängig von Protestantenanteil und Urbanisierungsgrad einen eigenständigen positiven (nicht-linearen) Einfluß auf das Stimmenergebnis der SPD hat.

Indirekte Effekte werden über die Multiplikation[13] zweier oder mehrerer Pfad- bzw. Regressionskoeffizienten errechnet, die einer Kausalkette zugeordnet sind, die zwischen prädeterminierter und endogener Variable über mindestens eine intervenierende Variable verläuft. In unserem Pfadmodell sind zwei indirekte Effekte eingezeichnet: von EV12 und URBAN auf PROZSPD, beide vermittelt über GEWLOG. Der indirekte Effekt von EV12 beträgt $\beta_5 \cdot \beta_4 = 0.379 \cdot 0.272 = 0.103$. Der indirekte Effekt von URBAN beträgt $\beta_6 \cdot \beta_4 = 0.512 \cdot 0.272 = 0.139$. Der **totale kausale Effekt** ergibt sich aus der Addition von direktem und indirektem Effekt: 0.492 + 0.103 = 0.595 für EV12 und 0.441 + 0.139 = 0.580 für URBAN. Die totalen Effekte von URBAN und EV12 auf PROZSPD unterscheiden sich also in noch geringerem Maße als ihre direkten Effekte.

In der Forschungspraxis gibt es aber durchaus Fälle, in denen die Rangfolge der totalen Effekte von der Rangfolge der direkten Effekte abweicht. Dennoch werden in der einschlägigen Literatur häufig (einfache) Regressionsanalysen präsentiert, in denen offenkundige kausale Beziehungen zwischen den Regressorvariablen nicht modelliert sind. Das heißt, es werden nur die direkten Effekte ermittelt, die indirekten und damit die totalen Effekte werden unterschlagen. Solange sich Autor und Leser über diesen Sachverhalt im klaren sind, ist dagegen nicht unbedingt etwas einzuwenden. Häufig jedoch werden in solchen Fällen die direkten Effekte wie totale Effekte behandelt. Das heißt, es werden auf der Basis der geschätzten direkten Effekte Aussagen gemacht über die kausale Bedeutung der einzelnen Regressorvariablen, ohne daß die möglicherweise vorhandenen, aber nicht ermittelten indirekten Effekte berücksichtigt würden. Die in-

---

[13] Eine theoretische Begründung für diese Rechenregel findet sich beispielsweise in Asher (1983) oder in Opp/Schmidt (1976).

direkten Effekte sind aber kausaltheoretisch nicht weniger wichtig als die direkten Effekte. Wenn z.B allgemeine Strukturvariablen (wie der sozio-ökonomische Status) und von ihnen beeinflußte intervenierende Variablen (z. B. Wertvorstellungen) als gleichgeordnete »unabhängige« Variablen in einem (einfachen) Regressionsmodell (»Eingleichungssystem«) auftauchen, mit dem ein Verhaltensindikator (als abhängige Variable) erklärt werden soll, führt das fast zwangsläufig zu täuschenden Ergebnissen: die Rolle der Strukturvariablen wird bei ausschließlicher Betrachtung der direkten Effekte unvollständig dargestellt. (Vergl. in Teil I, Kap. 5 die Erörterung der verschiedenen Kausalmodelle, vor allem des Unterschieds zwischen »Scheinkausalität« und »Intervention«.)

Wir wollen unser Pfadmodell nun etwas erweitern, indem wir in die Gleichung (11-48 a) den multiplikativen Ausdruck für die Interaktion von EV12 (X_1) und URBAN (X_2) hinsichtlich ihrer Wirkung auf PROZSPD einfügen. (Eine interaktive Wirkung von EV12 und URBAN auf GEWLOG besteht nicht.) Für die **standardisierten** Variablen[14] erhalten wir somit folgendes Gleichungssystem:

$$(11\text{-}49\ a) \qquad Z_y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \beta_4 Z_4 + u_1$$

$$(11\text{-}49\ b) \qquad Z_4 = \beta_5 Z_1 + \beta_6 Z_2 + u_2$$

Zur Notation siehe die Erläuterungen zu den Gleichungen (11-48a und b).

Die geschätzten Beta-Koeffizienten sind in das Pfaddiagramm der Abb. 11.11 eingetragen.

Die direkten Effekte von EV12 und URBAN auf GEWLOG (die in diesem Falle gleich den totalen Effekten sind) ändern sich nicht, da Gleichung (11-49 b) identisch ist mit Gleichung (11-48 b). Der direkte Effekt von GEWLOG auf PROZSPD verändert sich in (11-49 a) nur geringfügig gegenüber (11-48 a). Für den Einfluß von EV12 und URBAN auf

---

[14] Für Modelle mit multiplikativen Komponenten können Beta-Koeffizienten **nicht**, wie in Abschnitt 11.3 für additive Modelle angegeben, aus den unstandardisierten Regressionskoeffizienten gemäß $\beta = b \cdot (s_x/s_y)$ berechnet werden. Um die Beta-Koeffizienten zu erhalten, müssen alle Variablen (einschließlich der log-transformierten) zuvor in ihre standardisierten Werte $z = (x-\bar{x})/s_x$ transformiert und in dieser Form in die Regressionsanalyse eingeführt werden. Der multiplikative Term als Produkt zweier standardisierter Variablen ist selbst nicht standardisiert. Signifikanztests bleiben aber von dieser Maßnahme unberührt (siehe Marsden 1981, S. 115).

PROZSPD müssen nun aber sämtliche Effekte als **bedingte** Effekte berechnet werden. In den meisten Einführungstexten zur Pfadanalyse wird dieses Problem übergangen. Dabei ist seine Lösung recht einfach. Wie die bedingten direkten Effekte bei interaktiven Beziehungen zu ermitteln sind, haben wir in Abschn. 11.4 ausführlich besprochen. Das gleiche Verfahren ist auch hier anzuwenden: Die bedingten direkten Effekte werden errechnet, indem man die partiellen Ableitungen zur Gleichung (11-49 a) bildet:

$$\text{(11-50 a)} \quad \frac{\partial Z_y}{\partial Z_1} = \beta_1 + \beta_3 Z_2 = 0{,}521 + 0{,}178 Z_2$$

$$\text{(11-50 b)} \quad \frac{\partial Z_y}{\partial Z_2} = \beta_2 + \beta_3 Z_1 = 0{,}466 + 0{,}178 Z_1$$

$$\text{(11-50 c)} \quad \frac{\partial Z_y}{\partial Z_4} = \beta_4 = 0{,}275$$

Dies entspricht den Gleichungen (11-31) und (11-32) in Abschn. 11.4. Bei der Interpretation ist zu beachten, daß die arithmetischen Mittel von Z_1 und Z_2 gleich null sind und daß die standardisierten Variablen auch negative Werte annehmen.

Einem Vorschlag von Stolzenberg (1979) folgend, lassen sich die totalen Effekte in gleicher Weise berechnen, wenn man zuvor das Gleichungssystem (11-49 a u. b) in seine sog. »reduzierte Form« bringt. Ausgangspunkt ist die Gleichung für diejenige endogene Variable, auf die die zu errechnenden Effekte gerichtet sind, hier also Gleichung (11-49 a) für den SPD-Stimmenanteil. Kommen in dieser Gleichung Variablen vor, die zwischen den prädeterminierten Variablen (deren Effekte zu ermitteln sind) und der abhängigen Variablen intervenieren, so muß diese Variable (in unserem Beispiel GEWLOG) durch ihre gewichteten prädeterminierten Variablen ersetzt werden. Man sucht also im Gleichungssystem die Strukturgleichung(en) der intervenierenden Variable(n), hier (11-49 b), und setzt die Ausdrücke der rechten Gleichungsseite anstelle der intervenierenden Variable(n) in die Ausgangsgleichung ein:

(11-51)

$$\begin{aligned} Z_y &= \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \beta_4(\beta_5 Z_1 + \beta_6 Z_2 + u_2) + u_1 \\ &= (\beta_1 + \beta_4\beta_5) Z_1 + (\beta_2 + \beta_4\beta_6) Z_2 + \beta_3 Z_1 Z_2 + \beta_4 u_2 + u_1 \end{aligned}$$

Die totalen Effekte der prädeterminierten Variablen erhält man sodann wiederum über die entsprechenden partiellen Ableitungen:

$$(11\text{-}52\ a)\quad \frac{\partial Z_y}{\partial Z_1} = \beta_1 + \beta_4\beta_5 + \beta_3 Z_2 = 0{,}521 + 0{,}275 \cdot 0{,}379 + 0{,}178 Z_2$$

$$= 0{,}662 + 0{,}178 Z_2$$

$$(11\text{-}52\ b)\quad \frac{\partial Z_y}{\partial Z_2} = \beta_2 + \beta_4\beta_6 + \beta_3 Z_1 = 0{,}466 + 0{,}275 \cdot 0{,}512 + 0{,}178 Z_1$$

$$= 0{,}607 + 0{,}178 Z_1$$

Der totale Effekt von $\log X_1$ (GEWLOG) auf PROZSPD(Y) ist identisch mit dem direkten Effekt von GEWLOG; siehe die Gleichung (11-50 c). Die totalen Effekte von EV12(X_1) und URBAN(X_2) lassen sich wegen ihrer Interaktion nur als bedingte Effekte angeben. Für deskriptive Zwecke können sie am ehesten miteinander verglichen werden, wenn man als Bedingung jeweils die arithmetischen Mittel von X_1 und X_2 in die entsprechenden Gleichungen einsetzt. Da aber die arithmetischen Mittel der z-standardisierten Variablen null sind, erübrigt sich die Rechenarbeit, und wir können die unter dieser Bedingung erzielten Effekte von X_1 (0.662) und X_2 (0.607) direkt aus (11-52a) und (11-52b) ablesen.

Die indirekten Effekte ergeben sich durch Subtraktion der direkten Effekte von den totalen Effekten der gleichen Variablen:
Der indirekte Effekt von X_1 auf Y ist

$$(11\text{-}53a)\quad \beta_1 + \beta_4\beta_5 + \beta_3 Z_2 - (\beta_1 + \beta_3 Z_2) = \beta_4\beta_5 = 0{,}104$$

Der indirekte Effekt von X_2 auf Y ist

$$(11\text{-}53b)\quad \beta_2 + \beta_4\beta_6 + \beta_3 Z_1 - \beta_2 + \beta_3 Z_1) = \beta_4\beta_6 = 0{,}141$$

Somit werden auch bei dieser Methode die indirekten Effekte ermittelt, indem man die einzelnen Pfadkoeffizienten der Kausalkette miteinander multipliziert.

Mit der von Stolzenberg vorgeschlagenen Methode lassen sich die kausalen Effekte auch auf der Basis der nicht-standardisierten Regressionskoeffizienten berechnen. Dabei stellt sich heraus, daß die totalen Effekte des Protestantenanteils und des Urbanisierungsgrades auf den SPD-Stimmenanteil unabhängig vom jeweils erreichten Niveau des gewerkschaftlichen Organisationsgrades angegeben werden können.

Die nicht-standardisierten Regressions- bzw. Pfadkoeffizienten schätzen wir nach der OLS-Methode über folgendes Gleichungssystem:

$$(11\text{-}54\ a)\quad Y = a_1 + b_1X_1 + b_2X_2 + b_3X_1X_2 + b_4(\log X_4) + e_1$$

$$(11\text{-}54\ b)\quad \log X_4 = a_2 + b_5X_1 + b_6X_2 + e_2$$

Die geschätzten b-Koeffizienten sind in das Pfaddiagramm der Abb. 11.12 eingetragen. Damit lassen sich die direkten Effekte von EV12, URBAN und GEWLOG auf PROZSPD wie gewohnt mit Hilfe der ersten Ableitungen bestimmen:

$$(11\text{-}55\ a)\quad \frac{\partial Y}{\partial X_1} = b_1 + b_3X_2$$

$$(11\text{-}55\ b)\quad \frac{\partial Y}{\partial X_2} = b_2 + b_3X_2$$

$$(11\text{-}55\ c)\quad \frac{\partial Y}{\partial(\log X_4)} = b_4$$ [10]

Um die totalen Effekte zu berechnen, ersetzen wir zunächst wieder die intervenierende Variable $\log X_4$ in Gleichung (11-54 a) durch die rechte Seite von (11-54 b):

$$(11\text{-}56)\quad Y = a_1+b_1X_1+b_2X_2+b_3X_1X_2+b_4(a_2+b_5X_1+b_6X_2+e_2)+e_1$$

$$= a_1+b_4a_2+(b_1+b_4b_5)X_1+(b_2+b_4b_6)X_2+b_3X_1X_2+b_4e_2+e_1$$

Die (wegen der Interaktion von X_1 und X_2) bedingten totalen Effekte von X_1 und X_2 erhalten wir über die partiellen Ableitungen:

$$(11\text{-}57\ a)\quad \frac{\partial Y}{\partial X_1} = b_1 + b_4b_5 + b_3X_2$$

$$= 0{,}042 + 7{,}609 \cdot 0{,}0067 + 0{,}0033 \cdot X_2$$

$$= 0{,}093 + 0{,}0033 \cdot X_2$$

$$(11\text{-}57\ b)\quad \frac{\partial Y}{\partial X_2} = b_2 + b_4b_6 + b_3X_1$$

$$= 0{,}0604 + 7{,}609 \cdot 0{,}011 + 0{,}0033 \cdot X_1$$

$$= 0{,}144 + 0{,}0033 \cdot X_1$$

Als spezifizierende Bedingung für den totalen Effekt von X_1 (X_2) tritt lediglich X_1 (X_2), nicht auch GEWLOG auf. Es sei dem Leser überlassen, die totalen Effekte unter verschiedenen Bedingungen zu berechnen, z. B. unter der, daß $X_1 = \bar{x}_1 = 65.9$ und $X_2 = \bar{x}_2 = 64.4$[15].

Sowohl X_1 als auch X_2 wirken über $\log X_4$ auch indirekt auf Y. Diese indirekten Effekte erhalten wir durch Subtraktion der direkten Effekte (siehe Gleichungen (11-55 a und b)) von den totalen:

(11-58 a) Indirekter Effekt von X_1 $= b_1+b_4b_5+b_3X_2-(b_1+b_3X_2)$
$= b_4b_5$
$= 0,051$

(11-58 b) Indirekter Effekt von X_2 $= b_2+b_4b_6+b_3X_1-(b_2+b_3X_1)$
$= b_4b_6$
$= 0,084$

Die indirekten Effekte von EV12 und URBAN auf PROZSPD sind also sehr gering. Es mag auf den ersten Blick überraschen, daß sie konstant sind, daß der Effekt von X_1 bzw. X_2 nicht durch das jeweils erreichte Niveau des gewerkschaftlichen Organisationsgrades modifiziert wird. Diese Spezifikation bleibt deshalb aus, weil die logarithmierte Variable (GEWLOG) in der Kausalkette einmal als abhängige und einmal als unabhängige Variable auftritt, die nicht-linearen Komponenten sich somit wechselseitig aufheben.

Zum Schluß sei noch darauf hingewiesen, daß alle diese Berechnungen unter der Voraussetzung stehen, daß das Pfadmodell als Modell kausaler Beziehungen korrekt spezifiziert wurde. Unter bestimmten Bedingungen läßt sich die Angemessenheit des Modells anhand der Beobachtungsdaten testen. In diesem Zusammenhang ist es nötig, nicht nur die totalen **kausalen** Effekte in direkte und indirekte zu zerlegen, sondern auch die bivariaten Korrelationen soweit wie möglich in kausale, (bloß) korrelierte und Restgrößen zu zerlegen. Dies ist ein wichtiges Anliegen der Pfadanalyse, auf das wir in diesem knappen Überblick aber nicht eingehen wollen (siehe z. B. Opp/Schmidt 1976, S. 152).

---

[15] Die Mittelwerte haben sich gegenüber denen in Abschnitt 11.4 leicht verändert, da, wie oben erklärt, 88 Fälle aus der Stichprobe entfernt wurden.

Abb. 11.1: Kausalschema einer additiven Beziehung

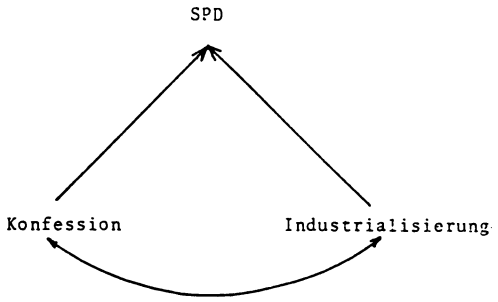
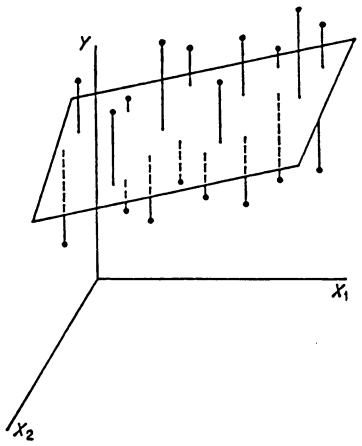


Abb. 11.2: Geometrische Darstellung einer Regression mit zwei unabh. Variablen



Quelle: Blalock 1960, S. 329

Abb. 11.3: SPSS^X-Ergebnisausdruck (Auszug) zur multiplen Regression mit zwei unabhängigen Variablen

* * * * M U L T I P L E R E G R E S S I O N * * * *

Beginning Block Number 1. Method: Enter EV12 INDUSTRY

Variable(s) Entered on Step Number 1.. EV12 PROTESTANTENANTEIL
2.. INDUSTRY

Multiple R	.79345
R Square	.62956
Adjusted R Square	.62767
Standard Error	10.80533

Analysis of Variance

	DF	Sum of Squares	Mean Squar
Regression	2	77782.11855	38891.0592
Residual	392	45768.06035	116.7552

F = 333.09900 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
EV12	.273146	.016593	.509425	16.461	.0000
INDUSTRY	.644875	.036121	.552496	17.853	.0000
(Constant)	-15.717589	1.861533		-8.443	.0000

Abb. 11.4: Residuenplots als Ergänzung zu Abb. 11.3

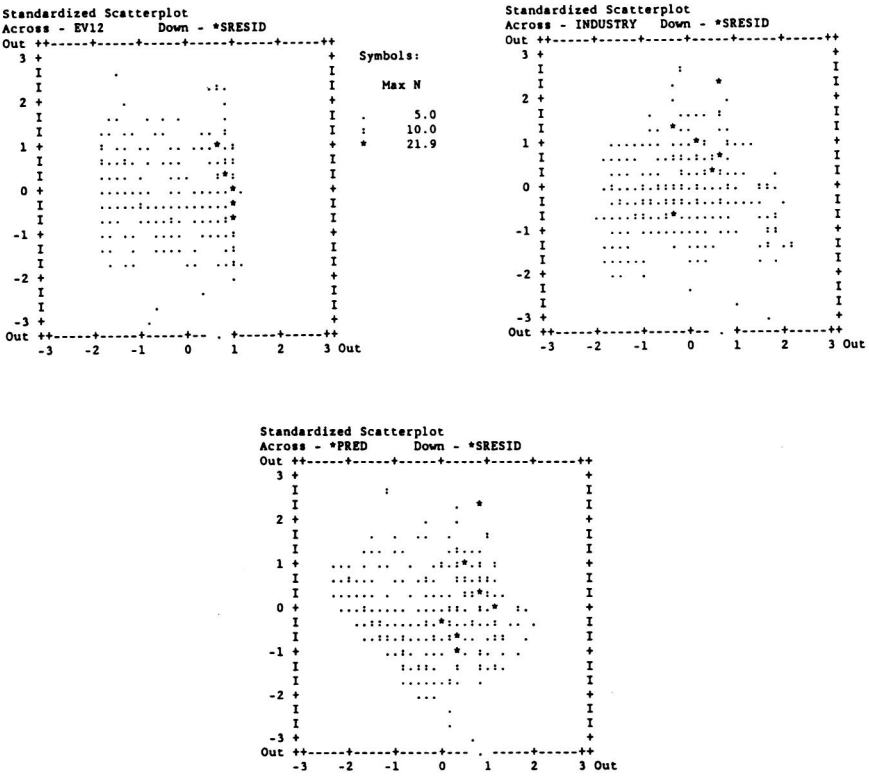


Abb. 11.5: Deskriptive Statistiken und Ergebnisausdruck zu einer zweistufigen Regression mit drei unabh. Variablen

	Mean	Std Dev	Label
PROZSPD	29.476	17.708	
INDUSTRY	43.316	15.171	
EV12	63.188	33.026	
URBAN	60.915	28.222	

N of Cases = 395

Correlation:

	PROZSPD	INDUSTRY	EV12	URBAN
PROZSPD	1.000	.611	.573	.666
INDUSTRY	.611	1.000	.115	.749
EV12	.573	.115	1.000	.089
URBAN	.666	.749	.089	1.000

Variable(s) Entered on Step Number 1.. EV12 PROTESTANTENANTEIL
2.. INDUSTRY

Multiple R	.79345			Analysis of Variance			
R Square	.62956	R Square Change	.62956		DF	Sum of Squares	Mean Squar
Adjusted R Square	.62767	F Change	333.09900	Regression	2	77782.11855	38891.0592
Standard Error	10.80533	Signif F Change	.0000	Residual	392	45768.06035	116.7552
				F = 333.09900		Signif F = .0000	

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance Above: Correlation

	EV12	INDUSTRY
EV12	2.753E-04	-.11510
INDUSTRY	-6.899E-05	.00130

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
EV12	.273146	.016593	.509425	16.461	.0000
INDUSTRY	.644875	.036121	.552496	17.853	.0000
(Constant)	-15.717589	1.861533		-8.443	.0000

Forts. Abb. 11.5

Beginning Block Number 2. Method: Enter URBAN

Variable(s) Entered on Step Number 3.. URBAN

Multiple R	.85297		
R Square	.72756	R Square Change	.09800
Adjusted R Square	.72546	F Change	140.63944
Standard Error	9.27839	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Squar
Regression	3	89889.56176	29963.1872
Residual	391	33660.61714	86.0885

F = 348.05084 Signif F = .0000

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance Above: Correlation

	EV12	INDUSTRY	URBAN
EV12	2.030E-04	-.07370	-.00377
INDUSTRY	-4.899E-05	.00218	-.74700
URBAN	-1.345E-06	-8.717E-04	6.256E-04

---------- Variables in the Equation ----------

VARIABLE	B	SE B	BEta	T	Sig T
EV12	.272508	.014248	.508236	19.126	.0000
INDUSTRY	.231583	.046654	.198408	4.964	.0000
URBAN	.296616	.025012	.472729	11.859	.0000
(Constant)	-15.843273	1.598507		-9.911	.0000

Abb. 11.6: Ergebnisausdruck zur multiplen Regression, additives und multiplikatives Modell

Correlation:

	PROZSPD	INDUSTRY	EV12	URBAN	URBEV12
PROZSPD	1.000	.611	.573	.666	.857
INDUSTRY	.611	1.000	.115	.749	.501
EV12	.573	.115	1.000	.089	.715
URBAN	.666	.749	.089	1.000	.678
URBEV12	.857	.501	.715	.678	1.000

Beginning Block Number 1. Method: Enter EV12 URBAN

Variable(s) Entered on Step Number 1.. URBAN
2.. EV12

Multiple R	.84284		
R Square	.71039	R Square Change	.71039
Adjusted R Square	.70891	F Change	480.76305
Standard Error	9.55407	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Squar
Regression	2	87768.32710	43884.1635
Residual	392	35781.85181	91.2802

F = 480.76305 Signif F = .0000

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance Above: Correlation

	URBAN	EV12
URBAN	2.932E-04	-.08873
EV12	-2.223E-05	2.141E-04

Forts. Abb. 11.6

---------- Variables in the Equation ----------

VARIABLE	B	SE B	BETA	T	SIG T
URBAN	.389359	.017122	.620538	22.740	.0000
EV12	.277721	.014632	.517958	18.981	.0000
(Constant)	-11.790773	1.415141		-8.332	.0000

---------- Variables not in the Equation ----------

Variable	Beta In	Partial	Min Toler	T	Sig T
URBEV12	.613401	.374681	.108057	7.991	.0000

Beginning Block Number 2. Method: Enter URBEV12

Variable(s) Entered on Step Number 3.. URBEV12

Multiple R	.86663		
R Square	.75104	R Square Change	.04066
Adjusted R Square	.74913	F Change	63.85507
Standard Error	8.86941	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Squar
Regression	3	92791.58165	30930.5272
Residual	391	30758.59725	78.6664

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance Above: Correlation

	URBAN	EV12	URBEV12
URBAN	.00114	.77070	-.88239
EV12	7.913E-04	9.237E-04	-.89457
URBEV12	-1.493E-05	-1.362E-05	2.508E-07

---------- Variables in the Equation ----------

VARIABLE	B	SE B	BETA	T	SIG T
URBAN	.151154	.033783	.240900	4.474	.0000
EV12	.060463	.030392	.112766	1.989	.0473
URBEV12	.004002	5.0080E-04	.613401	7.991	.0000
(Constant)	.713958	2.043205		.349	.7270

Abb. 11.7: Nicht-linearer Zusammenhang zwischen SPD-Stimmenanteil und Anteil an Gewerkschaftsmitgliedern

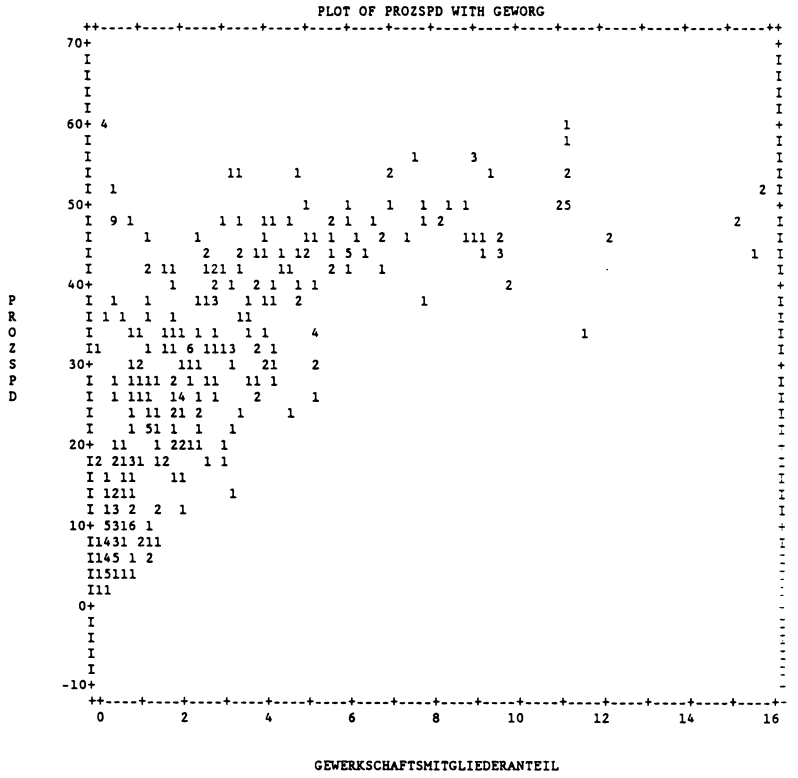


Abb. 11.8: Ergebnisausdruck zur multiplen Regression unter Einbeziehung des gewerkschaftlichen Organisationsgrades

EQUATION NUMBER 1 DEPENDENT VARIABLE.. PROZSPD

Multiple R	.85639		
R Square	.73341	R Square Change	.73341
Adjusted R Square	.72990	F Change	209.07959
Standard Error	8.07479	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Squar
Regression	4	54529.75969	13632.4399
Residual	304	19821.45557	65.2021

F = 209.07959 Signif F = .0000

Var-Covar Matrix of Regression Coefficients (B)
Below Diagonal: Covariance Above: Correlation

	URBEV12	GEWLOG	URBAN	EV12
URBEV12	3.176E-07	.01385	-.85285	-.90311
GEWLOG	8.177E-06	1.09726	-.29736	-.19954
URBAN	-1.947E-05	-.01262	.00164	.83167
EV12	-1.913E-05	-.00786	.00127	.00141

---------- Variables in the Equation ----------

VARIABLE	B	SE B	BETA	T	SIG T
URBEV12	.003310	5.6360E-04	.544066	5.872	.0000
GEWLOG	7.609342	1.047503	.274864	7.264	.0000
URBAN	.060375	.040514	.101065	1.490	.1372
EV12	.041959	.037592	.085611	1.116	.2652
(Constant)	8.968982	2.783823		3.222	.0014

Abb. 11.9: Pfaddiagramm (rekursives Pfadmodell)

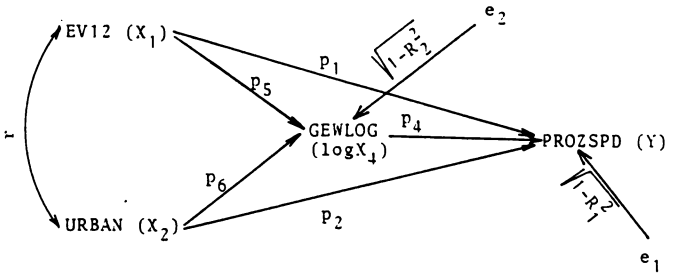


Abb. 11.10: Rekursives Pfadmodell ohne multiplikativen Term mit standardisierten Regressionskoeffizienten

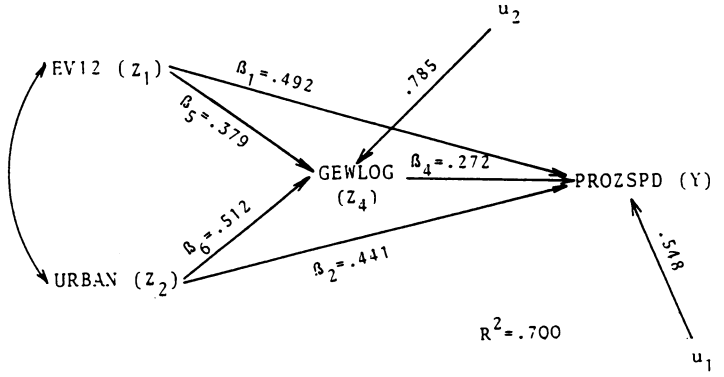


Abb. 11.11: Pfadmodell mit multiplikativem Term und standardisierten Regressionskoeffizienten

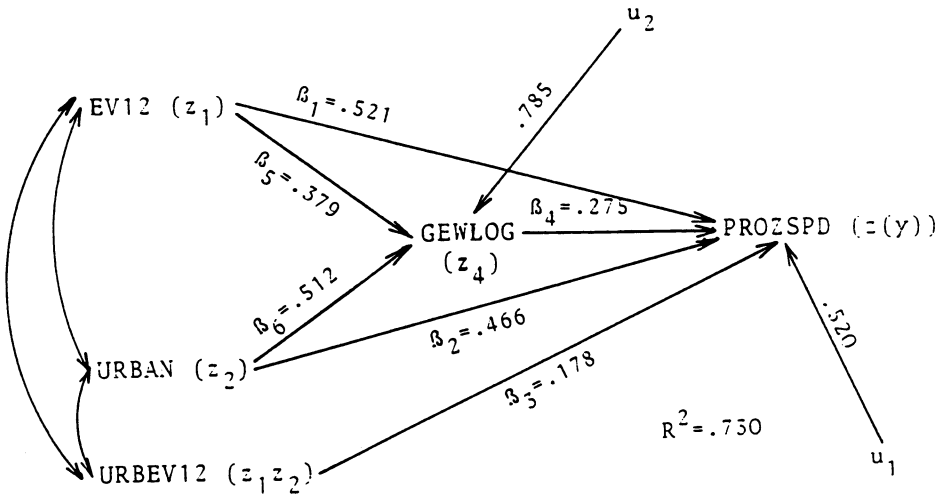
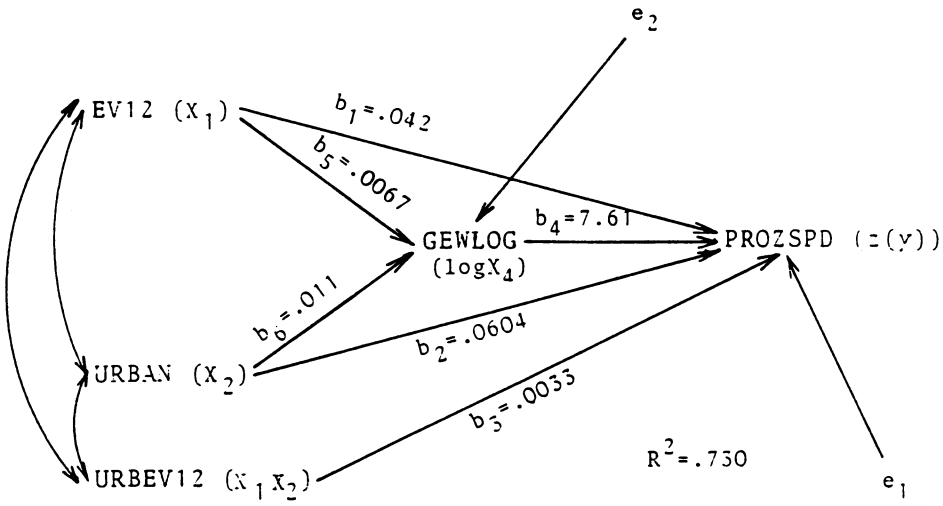


Abb. 11.12: Pfadmodell mit multiplikativem Term und nicht-standardisierten Regressionskoeff.



KAPITEL 12

Nicht-lineare Regression

12.1 Linearisierung von Beziehungen

Wie wir in Kap. 11.5 sahen, ist die Beziehung zwischen zwei Variablen nicht immer linear im Sinne der Gleichung $E(Y)=a + bX$. Sie kann auch kurvenförmig verlaufen. So nimmt z. B. in vielen Gesellschaften das Einkommen (Y) nach Beginn des Erwerbslebens in den ersten Jahrzehnten zu, fällt dann aber, wenn das Lebens- oder »Dienstalter« (X) einen bestimmten Schwellenwert überschritten hat, mehr oder weniger deutlich ab. Die Beziehungsform entspricht in diesem Falle schematisch einem umgekehrten U (s. Abb. 12.1).

Mathematisch kann eine solche Beziehung durch ein Polynom zweiter Ordnung ausgedrückt werden:

$$(12\text{-}1) \qquad \hat{Y} = a + b_1X + b_2X^2$$

Sollen die Parameter (a, b_1, b_2) anhand von Beobachtungsdaten geschätzt werden, kann man die Gleichung durch einen Fehlerterm erweitern und das gleiche Schätzverfahren (OLS, Kleinstquadratemethode) anwenden, das wir in den vorangegangenen Abschnitten besprochen haben:

$$(12\text{-}2) \qquad y_i = a + b_1x_i + b_2x_i^2 + e_i \;, \quad \Sigma e_i^2 = \min.$$

Technisch handelt es sich hierbei um eine multiple Regression (siehe Kap. 11), da X^2 im Schätzalgorithmus wie eine zweite unabhängige Variable behandelt wird, für die ein Koeffizient, b_2, zu schätzen ist. Inhaltlich bleibt es jedoch eine Beziehung zwischen den beiden Variablen Y (hier: Einkommen) und X (hier: Alter). Der Einfluß der unabhängigen Variablen drückt sich aber in zwei Regressionskoeffizienten aus. Formal ähnelt die Situation der einer interaktiven Beziehung; multipliziert werden aber nicht zwei unterschiedliche Variablen, sondern eine Variable »interagiert« gleichsam mit sich selbst. Wir werden gleich erläutern, wie die beiden b-Koeffizienten zu interpretieren sind, wollen zuvor aber noch den entsprechenden SPSS^x-Befehl zitieren. Er ändert sich gegenüber der einfachen Regression nur minimal: die Variablenliste muß lediglich um X^2 erweitert werden:

REGRESSION VARIABLES = Y, X, X^2
/DEPENDENT = Y / ENTER

Das setzt voraus, daß X^2 zuvor über ein COMPUTE-Statement

COMPUTE X^2 = X**2

gebildet worden ist.

Stolzenberg (1979) berichtet über eine US-amerikanische Untersuchung, in der solche Regressionsgleichungen für Arbeitnehmer aus verschiedenen Industriebranchen geschätzt wurden mit Y:= Stundenlohn in Dollar, X:= Lebensalter in Jahren. Für Elektriker wurden beispielsweise folgende Steigungskoeffizienten geschätzt:

$b_1 = 0.031$, $b_2 = -0.00032$.

Wie hat man diese beiden Koeffizienten zu interpretieren? Wäre $b_2 = 0$, läge eine lineare Beziehung vor, in der pro zusätzlichem Lebensjahr im Durchschnitt 3.1 Cents je Arbeitsstunde mehr verdient würden. Tatsächlich aber liegt eine nicht-lineare Beziehung vor, bei der das Produkt $0.00032 \cdot X^2$ von dem Produkt $0.031 \cdot X$ abgezogen werden muß. Da das negative Produkt wegen des quadrierten Alters bei zunehmendem X immer stärker ins Gewicht fällt, nimmt der Zuwachs in Y und damit die Steigung der Kurve mit zunehmendem X immer weiter ab; sie erreicht bei einem bestimmten Lebensalter ein Maximum und wird dann negativ.

Wir wissen aus der Schulmathematik, daß man die Steigung einer (stetigen) Kurve $Y = f(X)$ für jeden beliebigen Wert $X = x_i$ erhält, indem man die 1. Ableitung $\delta y/\delta x = y'$ bildet. Davon haben wir schon im vorigen Kapitel regen Gebrauch gemacht. In unserem Beispiel ist die erste Ableitung von (12-1)

$$\hat{Y}' = b_1 - 2b_2X \qquad (12\text{-}3)$$

Die Steigung ist also nicht mehr konstant (wie bei der linearen Beziehung $\hat{y} = a + bx$, $\hat{y}' = b$), sondern verändert sich (negativ) mit X. Wenn man nicht nur wissen will, wie groß die Steigung in jedem Kurvenpunkt ist, sondern wie groß der Y-Zuwachs beim Übergang von einem beliebigen Punkt x_1 zu einem anderen Punkt x_2 ist, braucht man nur die entsprechenden x-Werte in (12-1) einzusetzen und die Differenz $\hat{y}_2 - \hat{y}_1$ zu bilden. Wenn wir Gleichung (12-3) $\hat{y}'$ gleich Null setzen und nach x lösen, erhalten wir die Altersstufe, zu der ein maximaler Stundenverdienst erreicht wird (da im Maximum der Kurve die Steigung gleich Null sein muß):

(12-4) $0 = 0{,}0031 - 2 \cdot 0{,}00032x$

$x = 0{,}031/0{,}00064 = 48{,}44$

In der Mitte des 49. Lebensjahres beginnt also die Verdienstkurve der Elektriker zu fallen.

Ebenso sind Beziehungen denkbar, bei denen die Werte der Y-Variable mit steigenden X-Werten zunächst abnehmen, dann aber nach einem bestimmten $X = x_i$ allmählich ansteigen (U-förmige Beziehung). Auch dies entspricht wieder einem Polynom zweiter Ordnung gemäß Gleichung (12-1), in der die Koeffizienten aber umgekehrte Vorzeichen erhalten. Eine Auswahl von Beziehungsformen, die sich mit Parabelgleichungen darstellen lassen, gibt Abb. 12.2.

Polynome stellen Gleichungen dar, die zwar »in den Variablen« nicht-linear, »in den Regressionskoeffizienten« aber linear sind.

Deshalb können die Parameter weiterhin mit dem Kleinstquadrateverfahren geschätzt werden. Hyperbelfunktionen (siehe Abb. 12.3) stellen einen weiteren Typ nicht-linearer Beziehungen dar, die ebenso problemlos mit der Kleinstquadratemethode geschätzt werden können, nachdem man die X-Variable zu $1/X$ transformiert hat.
Gelegentlich werden Wurzeltransformationen angewandt:

(12-5) $Y = a + b\,X + e, \qquad \hat{Y} = a + bX^* \;,\; X^* = \sqrt{X}$

die ebenfalls »linear in den Parametern« sind. Mit ihnen kann man Beziehungen formalisieren, in denen die Regressionslinie zunächst steil, dann zunehmend flacher ansteigt. Das war auch der Ausgangspunkt für die logarithmierte Transformation der GEWORG-Variable, die wir in Kap. 11.5 angewandt haben. Gegenüber der Wurzeltransformation hat sie den Vorteil, daß der Steigungskoeffizient b direkt im Sinne einer Veränderungsrate interpretierbar ist. Immerhin kann auch nach einer Wurzeltransformation die »punktuelle« Steigung der Regressionskurve für jeden beliebigen Punkt der X-Achse wiederum über die 1. Ableitung von Gleichung (12-5) bestimmt werden.

(12-6) $\hat{Y}' = (b/2)/\sqrt{X}$

Wir kommen nun zu Funktionsgleichungen, die zunächst in den Parametern nicht linear sind, aber durch eine Transformation »linearisiert« werden können. Dazu gehört z. B. die folgende Exponentialfunktion:

(12-7) $Y = aX^b \cdot e$ oder $E(Y) = aX^b$

Je nach der Größe von b kann diese Funktion die in Abb. 12.4 dargestellte Verlaufsformen annehmen.

In der Ökonometrie ist der im Exponenten stehende Regressionskoeffizient b als **Elastizität** bekannt. Er gibt den prozentualen Zuwachs in Y an, wenn X um 1 Prozent wächst. Diesen Koeffizienten kann man wiederum mit der Kleinstquadratmethode schätzen, wenn (12-7) logarithmisch transformiert wird zu:

(12-8) $\log(Y) = a + b \cdot \log(X) + \log(e)$

Diese Gleichung ist nun in den Parametern wieder linear. Wie bereits in Abschnitt 11.5 erläutert, ist es im Prinzip gleichgültig, welche Logarithmus-Basis man wählt. Anschaulicher oder vertrauter ist wahrscheinlich die Basis zehn. Doch läßt sich die Differentialrechnung vereinfachen, wenn man als Basis die Eulersche Zahl e $\approx$ 2.718 wählt (sog. »natürlicher Logarithmus«, abgekürzt ln). Es gilt: $\log_e(x) \approx 2.303 \cdot \log_{10}(x)$.

Die logarithmierten Variablen werden wie gehabt in die Schätzgleichung und in das entsprechende SPSS^x-Kommando für die Regressionsanalyse eingesetzt.

Ein sozialwissenschaftliches Beispiel liefert Tufte (1974, S. 113 ff.), der den Zusammenhang zwischen der Bevölkerungsgröße (X) und der Größe der Parlamente (Y) in 29 Demokratien untersuchte. Die Beziehung zwischen den Originalvariablen sieht im Streudiagramm wie in Abb. 12.5 dargestellt, aus. In dieser Form ist die Beziehung praktisch nicht interpretierbar. Werden aber beide Größen logarithmiert stellt sie sich ganz anders dar (s. Abb. 12.6).

Zwischen einer **prozentualen** Zunahme der Bevölkerungsgröße und der **prozentualen** Zunahme der Parlamentsgröße besteht offensichtlich ein linearer Zusammenhang.

Wie man zu der Interpretation des b-Koeffizienten als »Elastizität« kommt, kann man sich klarmachen, indem man von Gleichung (12-7) nach Eliminierung des Fehlers die erste Ableitung bildet:

(12-9)

$$\hat{y} = ax^b$$

Unter Anwendung der Regeln für das Differenzieren von Potenzen wird daraus

$$\hat{y}' = \partial y / \partial x = a \cdot b \cdot x^{b-1}$$

Da $x^{b-1} = x^b/x$, gilt auch

$$\partial\hat{y}/\partial x = a \cdot b \cdot x^b/x = (ax^b/x)b$$

Im Zähler des Klammerausdrucks steht der Ausdruck der ersten Zeile, somit ist

$$\partial\hat{y}/\partial x = (\hat{y}/x)b$$

$$(\partial\hat{y})x/(\partial x)\hat{y} = b$$

$$(\partial\hat{y})/\hat{y}) \cdot (x/\partial x) = b$$

$$(\partial\hat{y})/\hat{y} = b \cdot \partial x/x$$

Der Regressionskoeffizient b drückt also das konstante Verhältnis zweier Veränderungsraten aus. Die Steigung der Regressionslinie verändert sich mit den X-Werten. Sie kann über die 1. Ableitung (siehe die zweite Zeile in (12-9)) für jeden beliebigen Punkt $X = x_i$ berechnet werden.

Unter Umständen sind aber auch die Veränderungsraten zweier Variablen nicht linear, sondern »quadratisch« aufeinander bezogen. Bei einer Ausdehnung der Analyse auf 135 Nationen fand Tufte folgende Beziehungen zwischen der Zahl der Abgeordneten (»members«) und der Bevölkerungsgröße (»P«) (siehe Abb. 12.7 a u. b).

Wie man beim Vergleich der beiden Determinationskoeffizienten r^2 sieht, führt das nicht-lineare Modell zu einer besseren »Anpassungsgüte«; es erklärt einen höheren Anteil der Y-Varianz. Unterstellt man das linearen Modell, erhält man eine konstante Elastizität von 0.396: Ein einprozentiger Bevölkerungszuwachs läßt die Parlamente im Schnitt um 0.4 % größer werden. Im nicht-linearen Modell

$$(12\text{-}10) \qquad \log\hat{Y} = 0{,}667 + 0{,}031(\log X)^2,$$

in dem die logarithmierten X-Werte quadriert wurden, läßt sich der mit zwei multiplizierte b-Koeffizient wiederum als Elastizität von Y bezüglich X interpretieren: Die 1. Ableitung von (12-10) führt zu dem Ausdruck

$$(12\text{-}11) \qquad (\partial\hat{Y}/\partial X)(1/\hat{Y}) = 2b(\log X)(1/X)$$

$$\frac{\partial\hat{Y}/\hat{Y}}{\partial X/X} = 2b(\log X) = 0{,}062(\log X)$$

Die Elastizität ändert sich also mit dem Logarithmus von X. Bei einer Bevölkerung von beispielsweise 10 Mill. ergibt sich eine Elastizität von 0.434, für eine Bevölkerung von 100 Mill. (Zunahme des Logarithmus um 1 Einheit) eine Elastizität von 0.496.

In anderen Fällen mag es angemessen sein, nicht beide, sondern nur eine der beiden Variablen zu logarithmieren und die jeweils andere mit den ursprünglichen Skalenwerten in die Regressionsgleichung einzusetzen. Wenn nur Y logarithmiert wird,

(12-12)

$$\ln Y = a + bX + u \quad \rightarrow \quad E(Y) = e^{(a+bX)}$$

(um den Fehlerterm von der Basis e ≈ 2,718 des natürlichen Logarithmus zu unterscheiden, verwenden wir hier den Buchstaben u für den Fehlerterm)

gibt der Steigungskoeffizient b **indirekt** an, um wieviel **Prozent** sich Y im Durchschnitt verändert, wenn sich das **Niveau** von X um eine Skaleneinheit erhöht. Diese Information erhält man durch folgende Umformung: (Log-Basis hoch b minus 1)· 100. Falls b klein ist (Daumenregel $b < 0.25$), ergibt sich die prozentuale Veränderung in Y je Einheitszuwachs in X näherungsweise mit b· 100 (siehe Tufte 1974, S. 125). Falls b positiv ist, erhält man eine zunehmend steiler werdende Kurve, wenn b negativ ist, eine zunehmend flacher werdende Kurve (siehe Abb. 12.8).

Wenn nur die **unabhängige** Variable zu einer beliebigen Basis c logarithmiert wird, gibt der Steigungskoeffizient b an, um wieviele Skaleneinheiten sich Y verändert, wenn X sich um den **Faktor** c vergrößert (d. h., wenn der Logarithmus um 1 Einheit zunimmt). Wählt man z. B. $c = 2$, gibt b den Y-Zuwachs an, falls X sich verdoppelt. Bei positivem Koeffizienten b erhält man wiederum (ähnlich wie bei der oben erläuterten Wurzelfunktion) eine Kurve, deren Steigung sich bei zunehmenden X-Werten abflacht. Ein Analysebeispiel hierzu haben wir schon in Kap. 11.5 erörtert. Weitere Analysebeispiele zu beiden Typen von Log-Transformationen findet man wiederum in Tufte (1974, S. 124)[1].

[1] Eine weitere Interpretationsmöglichkeit für das Modell, in dem nur die

Zum Schluß sei noch einmal daran erinnert, daß man Datentransformationen nicht nur zur Linearisierung von Beziehungen wählt. Sie können auch anderen Zwecken dienen, vor allem (wie wir schon in Abschn. 10.4 erwähnten) der Stabilisierung (dem »Konstantmachen«) von Varianzen (siehe Stoto/Emerson 1983; für weitere Zwecksetzungen siehe Schlittgen 1987, S. 159 ff.)

12.2 Nicht-lineare Regression und ML-Schätzung

In Kap. 10 wurde mehrmals darauf hingewiesen, daß Regressionsmodelle mit Prozentzahlen in der Regel nicht linear sind. Wir wollen diesen Einwand nun vertiefen und eine Alternative, ein nicht-lineares Regressionsmodell, erläutern. Zuvor ist noch ein anderer Problemkomplex zu erörtern, der stets bei der Analyse von Prozentzahlen zu bedenken ist, unabhängig davon, ob man ein lineares oder ein nicht-lineares Regressionsmodell schätzen möchte.

12.2.1 Zur Interpretation von dichotomisierten und prozentuierten Daten auf der Individual- und Aggregatebene

Prozentzahlen entstehen, das wurde ebenfalls schon in früheren Abschnitten erwähnt, aus dem Addieren, dem »Aggregieren« von Individualdaten. Im Wahlkreis Memel-Heydekrug zum Beispiel haben von $n_j = 22189$ Wahlberechtigten 3794 bei der Hauptwahl am 12. 1. 1912 für den Reichstagskandidaten der SPD votiert. Wenn man die Einzelstimmen y_{ij} ($i = 1,2, \ldots, n_j$) dieses hier mit »j« indizierten Wahlkreises mit »1« verkodet, falls sie für die SPD abgegeben wurden, und mit »0«, falls sie nicht der SPD zugute kamen, erhält man für diesen Wahlkreis ein arithmetisches Mittel von $3794/22189 = 0.171$. Multipliziert man diesen Anteilswert mit 100, wird daraus die Prozentgröße 17.1 %. Die 0/1-kodierte Variable der einzelnen Stimmen ist auf der sog. Individualebene definiert, Anteils-

---

unabhängige Variable logarithmiert wird, erhält man wiederum über die 1. Ableitung von $y = a + b(\ln x)$. Sie ist $\delta y/\delta x = b/x$. Daraus folgt $\delta y = b(\delta x/x)$. Somit erhält man bei einer sehr kleinen Änderungsrate in X (z. B. 1/100) den entsprechenden Änderungsbetrag in Y, indem man den Koeffizienten b mit der Änderungsrate von X multipliziert. Aus einer Zunahme von X um 1/100 folgt eine Veränderung in Y um b/100. Da diese Beziehung nur für sehr kleine Veränderungsraten gilt, folgt daraus im allgemeinen nicht, daß sich Y um den Betrag b verändert, falls sich X verdoppelt, d. h. um 100 % zunimmt (siehe Stolzenberg 1979, S. 470). Wenn nicht zur Basis e, sondern zur Basis a_e logarithmiert wird, gilt: $Y = a + b(\log_a X) \rightarrow Y' = b\,[(\log_a e)/X]$.

werte bzw. Prozentgrößen y_j ($j = 1,2, \ldots, n = 395$) sind auf der sog. Aggregatebene definiert, auf der nicht die einzelnen Wahlberechtigten, sondern die Wahlkreise die Untersuchungseinheiten bilden. Die Individualdaten sind in unserem Beispiel nicht bekannt; wir wissen nicht, welche Wähler in einem bestimmten Wahlkreis für den SPD-Kandidaten gestimmt haben und welche nicht. Uns liegen nur die Aggregatdaten für die Wahlkreise (»Kollektive«) vor. Man bezeichnet Anteilswerte als »analytische Kollektivmerkmale«, weil sie lediglich aus einer Rechenoperation (hier: Summieren und Dividieren), angewandt auf Individualdaten, hervorgehen. Auch die Variable »Industrialisierungsgrad« stellt in unserem Beispiel ein analytisches Kollektivmerkmal dar, denn es handelt sich dabei um den Prozentanteil aller Bewohner eines Wahlkreises, die in der Industrie beschäftigt sind (einschließlich der Haushaltsangehörigen).

Wenn sich Variablen sowohl als Individual- wie auch als (analytisches) Kollektivmerkmal interpretieren lassen, entsteht für den Datenanalytiker ein Problem, das wir schon in Kap. 11.4 gestreift hatten: Kann man von Variablenzusammenhängen, die man mit Aggregatdaten (mit anlytischen Kollektivmerkmalen) ermittelt hat, auf entsprechende Variablenzusammenhänge auf der Ebene der Individualdaten schließen? Wenn z. B. mit Wahlkreisdaten eine positive Korrelation zwischen dem Anteil der Beschäftigten in der Industrie einerseits und dem Stimmenanteil der SPD andererseits festgestellt wird, kann man allein daraus schon folgern, daß eine ebenso hohe Korrelation zwischen dem individuellen Beschäftigtsein in der Industrie und der individuellen Neigung, SPD zu wählen, besteht? Die Antwort ist: nein, dieser Schluß ist nicht zwingend; er kann falsch sein. In der Literatur wird dieses Problem unter dem Stichwort »ökologischer Fehlschluß« diskutiert (zur Einführung siehe Hummell 1972). Im allgemeinen lassen sich Korrelationskoeffizienten nicht von der Aggregat- auf die Individualebene übertragen. Wenn man z. B. in den USA festgestellt hat, daß die Kriminalitätsrate einzelner Stadtbezirke positiv mit dem Anteil der dort wohnenden schwarzen Bevölkerung korreliert, läßt sich daraus nicht folgern, daß Schwarze eher kriminell werden als Weiße. (Es könnten auch die Weißen zunehmend kriminell werden, wenn der Anteil der Schwarzen in ihrem Wohnbezirk zunimmt.) Unter sonst gleichen Bedingungen wird der Korrelationskoeffizient um so größer (im Absolutbetrag), je größer die Aggregateinheiten sind, je mehr Individuen zu einem Kollektiv zusammengefaßt wurden.

Wie aber läßt sich die Regressionsgerade im Falle einer dichotomen oder prozentualen abhängigen Variablen überhaupt interpretieren? Wir erinnern uns, daß die Punkte der Regressionsgeraden die bedingten Erwartungswerte (Mittelwerte) $E(Y|x_k)$ angeben (»k« indiziert hier einen bestimmten Wert der unabhängigen Variablen). Eine positive Beziehung zwischen einer 1/0-kodierten abhängigen Variablen und einer metrisch

gemessenen unabhängigen Variablen würde bedeuten, daß $Y = 1$ um so häufiger ($Y = 0$ also um so seltener) realisiert wird, je größer X. Das heißt: die mit »1« kodierte Merkmalsausprägung (in unserem Beispiel: das Votum für den SPD-Kandidaten) wird mit wachsendem X immer wahrscheinlicher. Die Erwartungswerte $E(Y \mid x_k)$ lassen sich als geschätzte Wahrscheinlichkeitsgrößen $\hat{\pi}_k = a + bx_k$ deuten, die sämtliche Werte innerhalb des Intervalls [0,1] annehmen können. (Auf die Begrenzung des Wertebereichs werden wir im nächsten Abschnitt näher eingehen.) Für den Anteilswert $p_j = \Sigma\, y_{ij}/n_j$, der aus einer Aggregierung der Individualdaten (der dichotomen y_{ij}-Werte im Wahlkreis j, $j = 1,2, \ldots, 395$) hervorgeht, gilt die gleiche Interpretation: Im ersten Falle erhält man den Erwartungswert $E(Y \mid x_k)$ als bedingten Mittelwert der 1/0-kodierten Individualdaten; im zweiten Falle erhält man $E(Y \mid x_k)$ als Mittelwert von Anteilswerten, die zuvor aus den gleichen (allerdings gruppierten) Individualdaten errechnet wurden. In beiden Fällen sind die Erwartungswerte als (geschätzte) Wahrscheinlichkeitsgrößen zu interpretieren.

In unserem Analysebeispiel sind uns sowohl bei der abhängigen Variable (Stimmabgabe für die SPD) wie auch bei der unabhängigen Variable (Beschäftigte in der Industrie) nur die Anteilswerte gegeben. Lägen die Individualdaten aus einer Stichprobe der einzelnen Wahlberechtigten vor, könnten wir sie in zweierlei Weise »verarbeiten«. Als erstes könnten wir feststellen, wie groß die Wahrscheinlichkeit ist, daß ein in der Industrie Beschäftigter (bzw. ein Mitglied seines Haushalts) im Unterschied zu einem außerhalb der Industrie Beschäftigten für die SPD votiert (hat). Die Antwort wäre einer einfachen 4-Felder-Tafel wie in Abb. 12.9 zu entnehmen.

Die durch 100 dividierte Prozentdifferenz zwischen den SPD-Stimmen in den beiden X-Gruppen gäbe die Differenz der Wahrscheinlichkeit an, mit der die in der Industrie Beschäftigten (bzw. deren Haushaltsmitglieder) im Unterschied zu den außerhalb der Industrie Beschäftigten die SPD wählen (bzw. gewählt haben). In dem fiktiven Beispiel aus Abb. 12.9 nimmt die Wahrscheinlichkeit von $\hat{\pi} = 0.1$ auf $\hat{\pi} = 0.6$ zu.

Als zweites können wir die Wahrscheinlichkeit schätzen, mit der irgendein Wahlberechtigter (unabhängig von seiner eigenen Beschäftigung innerhalb oder außerhalb der Industrie) für die SPD votiert (hat), und zwar in Abhängigkeit von dem in seinem Wahlkreis erreichten Anteil industriell Beschäftigter (einschließlich der Haushaltsangehörigen). Auch diejenigen Wähler, die selbst nicht in der Industrie beschäftigt sind, könnten eine stärkere Präferenz für die SPD entwickeln, wenn ihre »Umgebung« in wachsendem Maße durch die Industrialisierung geprägt wird. Man spricht in diesem Zusammenhang von einem »Kontexteffekt«. Wenn nur die Aggregatdaten (nur die Anteilswerte) bekannt sind, läßt sich keine Aussage darüber machen, in welchem Maße die beiden Personengruppen

(innerhalb und außerhalb der Industrie Beschäftigte) zu den SPD-Voten jeweils beigetragen haben. Die Regressionsparameter sind dann nur im Hinblick auf die Wahlkreise interpretierbar. Der Steigungskoeffizient beantwortet lediglich die Frage: In welchem Maße verändert sich der SPD-Stimmenanteil in Abhängigkeit vom Industrialisierungsgrad der Wahlkreise? Ein (vereinfachtes) Regressionsmodell (ein »lineares« Wahrscheinlichkeitsmodell) hierzu haben wir in Abschnitt 10.1 vorgestellt. Ein adäquateres Regressionsmodell für Anteilswerte wird in den nächsten Abschnitten erläutert.

Von Schätzproblemen abgesehen, läßt sich ein Regressionsmodell auch dann spezifizieren, wenn die abhängige Variable dichotom ist. Lägen die Daten über die einzelnen Stimmabgaben vor, könnten in einem Regressionsmodell sowohl Individual- als auch Kontexteffekt gleichzeitig berücksichtigt werden. Untersuchungseinheit wären die einzelnen Wahlberechtigten (von denen man eine Stichprobe gezogen hätte). Abhängige Variable Y wäre die 1/0-kodierte Stimmabgabe für (oder gegen) die SPD. Die erste unabhängige Variable X_1 wäre ein ebenfalls dichotomer Regressor, der angibt, ob der Wahlberechtigte in der Industrie beschäftigt ist oder nicht. Die zweite unabhängige Variable X_2 wäre beispielsweise ein multiplikativer Term, zusammengesetzt aus X und $\overline{X}$:= Anteil der in der Industrie Beschäftigten:

$$(12\text{-}13) \qquad Y_{ij} = a + b_1 X_{ij} + b_2 (X_{ij}\overline{X}) + e_{ij}$$

Das heißt, der Steigungskoeffizient von X wird im Sinne einer interaktiven Beziehung (siehe Abschn. 11.4) modifiziert durch den jeweils im Wahlkreis gegebenen Anteilswert $\overline{X}$ desselben Merkmals. Will man das Modell (12-13) auf die Aggregatebene »übertragen«, muß man die einzelnen Werte aller Individuen summieren und anschließend durch die Zahl der Fälle dividieren. Dann erhält man (siehe Jagodzinski/Weede 1981, S. 448)

$$(12\text{-}14) \qquad \overline{Y}_j = a + b_1 \overline{X}_j + b_2 \overline{X}_j^2 + \overline{e}_j$$

Die Gleichung enthält also nur noch Anteilswerte und aus dem multiplikativen Term ist ein quadratischer Ausdruck geworden.

12.2.2 Verletzung von Voraussetzungen im linearen Modell

Bisher haben wir nur lineare Regressionsmodelle behandelt, wobei die Linearität u. U. erst durch eine geeignete Variablentransformation zu erreichen war (siehe Abschn. 12.1). Bei dichotomen abhängigen Variablen oder entsprechenden Prozentgrößen kann, wie wir schon andeuteten, die

Linearitätsannahme aus theoretischen und formalen Gründen in Frage gestellt werden. In unserem Analysebeispiel wird man auf dieses Problem schon dadurch hingewiesen, daß im linearen Modell ein **negativer** Ordinatenabschnitt auftaucht (siehe Abschn. 10.1). Das heißt, es wird im Modell ein negativer Stimmenanteil der SPD für den Fall erwartet, daß der Industrialisierungsgrad den Wert 0 annimmt. In einem anderen Anwendungsbeispiel könnten sich auch Prognosewerte $\hat{y} > 1$ (bzw. größer 100 %) innerhalb des real möglichen Wertebereichs von X einstellen. Anteilswerte bzw. Wahrscheinlichkeiten außerhalb des Intervalls [0,1] sind jedoch unsinnig.

Das lineare Wahrscheinlichkeitsmodell enthält also durch die Begrenzung des Wertebereichs Parameterrestriktionen, die sich im bivariaten Falle aus zwei Ungleichungen ergeben:

$$(12\text{-}15) \qquad 0 \leq a + b \cdot X(min)$$

$$a + b \cdot X(max) \leq 1$$

In Einzelfällen können daraus theoretisch absurde Einschränkungen für die möglichen Werte des Steigungskoeffizienten folgen (siehe hierzu Näheres in Aldrich/Nelson 1984, S. 25).

Gegen die Linearitätsannahme sprechen unabhängig von formalen Gesichtspunkten oft auch substantielle Überlegungen. In unserem Beispiel etwa folgende: Die Mobilisierung der Wählerschaft für die SPD ist zunächst, solange der Anteil der industriell Beschäftigten niedrig ist, relativ schwach. Die wenigen, die in der Industrie arbeiten, sind selbst noch stark durch traditionelle Verhaltensmuster bestimmt. Steigende Zuwachsraten für die SPD lassen sich erst dann erwarten, wenn die Industrialisierung weiter fortgeschritten ist und sich eine durch sie geprägte Subkultur gegen die traditionelle Kultur erfolgreich abzugrenzen beginnt. Andererseits ist zu erwarten, daß sich die Zuwachsraten für die SPD nach einer Phase beschleunigter Mobilisierung wieder abschwächen. Es ist z. B. damit zu rechnen, daß sich bei voranschreitender Industrialisierung die politischen und ökonomischen Interessen weiter differenzieren und deshalb weniger gut durch eine einzelne politische Partei bündeln und vertreten lassen. Außerdem verstärkt der Erfolg einer Partei im allgemeinen die Anstrengungen der Konkurrenzparteien, ihrerseits Wähler für sich zu mobilisieren. Auf der anderen Seite läßt das Engagement des Durchschnittswählers für eine Partei tendenziell nach, wenn eine bestimmte Erfolgsschwelle überschritten ist. Diese inhaltlichen Erwägungen (die natürlich noch zu ergänzen wären) legen für den Zusammenhang zwischen Industrialisierung und Parteipräferenz ein Modell nahe, dessen funktionale Form einem gestreckten S entspricht (s. Abb. 12.10).

Mit entsprechenden mathematischen Funktionen beschäftigen wir uns im folgenden Abschnitt 12.3.3.

Zuvor ist noch auf eine weiteres Problem hinzuweisen, das bei der Regressionsanalyse mit Prozentdaten bzw. dichotomen abhängigen Variablen im linearen Modell auftritt. Wir wissen aus früheren Kapiteln (siehe insbesondere die Gleichungen (7-3), (7-4) und (8-42)), daß die Varianz (V) 1/0-kodierter Variablen oder entsprechender Anteilswerte von der Wahrscheinlichkeit π des mit »1« kodierten Ereignisses (hier: Stimmabgabe für die SPD) abhängt:

(12-16)

$$V(Y) = \pi(1-\pi) \quad \text{bei 1/0 kodierter Y-Variable}$$

$$V(Y) = \pi(1-\pi)/n \quad \text{bei Anteilswerten (wobei n = Stichprobenumfang)}$$

Die Fehlervarianz ist also abhängig von der jeweiligen Wahrscheinlichkeit $P(Y=1)$ für das interessierende Ereignis; bei Anteilswerten sind sie auch noch von der Menge n der Beobachtungen abhängig, für die der Anteil errechnet wurde.

Es läßt sich außerdem zeigen, daß die Fehlervarianz systematisch mit der unabhängigen Variablen korreliert (siehe Albrich/Nelson 1984, S. 13).

Die Fehlervarianz erreicht ein Maximum, $V(e_j)=0.25$, bei einem Anteilswert bzw. einer Wahrscheinlichkeit von $\pi = 0.5$ und verändert sich nur geringfügig in der Nähe dieses Wertes. Bei $\pi=0.3$ zum Beispiel ist das Produkt $\pi(1-\pi)=0{,}21$. Die Veränderungen werden aber um so größer, je stärker sich π den Werten 0 und 1 nähert. Das bedeutet, daß die Voraussetzung konstanter Fehlervarianzen (Homoskedastizität) bei Anteilswerten von $\pi \leqq 0{,}3$ oder $(1-\pi) \leqq 0{,}3$ (das ist eine Daumenregel) nicht ausreichend erfüllt ist, wenn man die Effizienz des Schätzers optimieren und ihre Standardfehler erwartungstreu schätzen will. Wie schon in Abschnitt 10.4 erwähnt, wird zur Behebung dieses Defektes in der Regel vorgeschlagen, die beobachteten Anteilswerte p_j mit dem reziproken Wert ihrer geschätzten Standardabweichung zu gewichten, laut Gleichung (12-16) also mit

$$\sqrt{\frac{n_j}{p_j(1-p_j)}} \quad , \; p_j = \hat{\pi}_j$$

Dabei würden die Extremwerte (also Anteile nahe 0 oder 1) am stärksten

gewichtet, weil der Nenner des Bruchs dort am geringsten ist. Die Frage ist, ob man das möchte. Aldrich/Nelson (1984, S. 29 f.) haben gegen dieses Verfahren eingewandt, daß gerade die Extremwerte besonders stark von der wahren Regressionslinie abweichen können, wenn die Beziehung nicht-linear ist (womit aus den oben genannten Gründen zu rechnen ist). Sie geben zu bedenken, daß Gewichtungsverfahren zur Varianzstabilisierung nur angemessen sind, wenn das Modell korrekt spezifiziert worden ist (siehe auch Hanushek/Jackson 1977, S. 182). Das im nächsten Abschnitt zu besprechende Verfahren der logistischen Regression (nicht notwendigerweise, aber in der Regel verbunden mit der Maxium-Liklihood-Methode) antwortet auf beide Problemlagen: es enthält eine angemessene implizite Gewichtung der Beobachtungsdaten (siehe hierzu Linder/Berchtold 1976, S. 41,59), und es befreit von der restriktiven Linearitätsannahme.

12.2.3 Logistische Regression

Im vorangegangenen Abschnitt haben wir sowohl formale wie auch inhaltliche Gründe genannt, die dafür sprechen, dichotome abhängige Variablen bzw. Anteilswerte oder Prozentzahlen nicht im Rahmen eines linearen Modells zu analysieren, sondern eine S-förmige Beziehung zwischen abhängiger und erklärender Variable anzunehmen. Die S-Kurve selbst mag unterschiedlich gestaltet sein; die beiden Enden können z. B. unterschiedlich stark gestreckt oder gestaucht werden; das S kann mehr oder weniger steil stehen. Außerdem kann es in zwei spiegeldbildliche Hälften geteilt sein oder auch nicht. Die jeweils gewünschte Form läßt sich durch unterschiedliche mathematische Funktionen realisieren, die in der Literatur bestimmte Etiketten erhalten haben (siehe Abb 12.11).

Neben der sog. »Probit«-Funktion ist die »logistische« Funktion wohl diejenige, die am häufigsten zur Spezifikation nicht-linearer Regressionsmodelle herangezogen wird. Ihre formalen Eigenschaften wollen wir nun etwas näher betrachten. Da lediglich die Form der Regressionsbeziehung interessiert, eliminieren wir aus den folgenden Gleichungen den Fehlerausdruck e_i. Als abhängige Variable betrachten wir also den bedingten Erwartungswert $E(Y \mid x_i)$.

Dabei soll Y eine dichotome (1/0-kodierte) oder in Anteilswerten vorliegende abhängige Variable darstellen. $P = \hat{\pi}$ steht in beiden Fällen für die geschätzte Wahrscheinlichkeit, mit der das interessierende Merkmal oder Ereignis (in unserem Beispiel die Stimmabgabe für die SPD) eintritt. Die logistische Funktion ist dann wie folgt definiert:

$$(12\text{-}17) \qquad P_i = \frac{e^{\alpha+\beta x(i)}}{1+e^{\alpha+\beta x(i)}} = \frac{1}{1+e^{-(\alpha+\beta x(i))}}$$

X bezeichnet die unabhängige Variable des Modells. Wir gehen zunächst von einer einzigen Regressorvariable aus.

Um die Notation zu vereinfachen definieren wir im folgenden

$$(12\text{-}18) \qquad z := \alpha + \beta x \quad \Rightarrow \quad x = (z-\alpha)/\beta$$

Somit läßt sich Gleichung (12-17) wie folgt umformen (den Index i sparen wir uns):

$$\begin{aligned} (12\text{-}19) \qquad P(1 + e^z) &= e^z \\ P &= e^z - e^z P \\ &= e^z(1 - P) \\ P/(1-P) &= e^z \\ \log[P/(1-P)] &= z = \alpha + \beta x \end{aligned}$$

Der Ausdruck auf der linken Seite in der letzten Zeile von (12-19) wird als »Logit« bezeichnet. Der Logit-Wert ist also (in diesem Modell) **linear** abhängig von der Regressorvariablen. Falls Y eine dichotome Variable ist und die Daten nicht gruppiert sind, ist die Größe P/(1-P) nicht bekannt und die Parameter α und β können nicht über die Gleichung (12-19) mit Hilfe der Kleinstquadratmethode direkt geschätzt werden. Wenn die Daten nach Kategorien der unabhängigen Variablen gruppiert sind, kann man die Wahrscheinlichkeiten über die relativen Häufigkeiten schätzen, mit denen die einzelnen Kategorien besetzt sind. Unter Berücksichtigung der nicht-konstanten Varianzen lassen sich die Modellparameter sodann über ein gewichtetes Kleinstquadrateverfahren schätzen (siehe Hanushek/Jackson 1977, S. 190 ff.). Das gleiche gilt, wenn von vornherein nur Anteilswerte (also Aggregatdaten) vorliegen. Wir werden jedoch in Abschn. 10.2.4 ein allgemeineres, auch auf Individualdaten anwendbares Schätzverfahren kennenlernen. Zuvor sollen aber noch einige formale Eigenschaften der Logitfunktion erläutert werden.

Offensichtlich impliziert der Ausdruck (12-17), daß P_i nur Werte größer O und kleiner 1 annehmen kann, ohne daß der Wertebereich von X (bzw. z) eingeschränkt wird. (Beim linearen Modell konnten Wahrscheinlich-

keitswerte außerhalb des Intervalls [0,1] auftreten.). Wie leicht nachgerechnet werden kann, nähert sich P_i dem Wert 1, wenn z (also das Logit) ins positiv Unendliche wächst, und es nähert sich dem Wert 0 in dem Maße, wie sich z gegen negativ unendlich entwickelt. So entsprechen z. B. den Anteilswerten $P_1 = 0.01$, $P_2 = 0.99$ die Logits $z_1 = -4.6$ und $z_2 = 4.6$.

Wie Abbildung 12.12 zeigt, verläuft die in (12-17) ausgedrückte S-Kurve symmetrisch um den Punkt ($z_i = 0$, $P_i = 0.5$), in dem sie ihre größte Steigung hat.

Die den z-Werten entsprechenden X-Werte ergeben sich aus der Definitionsgleichung (12-18).

Für die inhaltliche Interpretation des logistischen Modells ist ein weiterer formaler Aspekt interessant: Die 1. Ableitung der Funktion (12-17) in Verbindung mit (12-18) nach z

$$(12\text{-}20) \qquad \partial P/\partial z = P(1-P)$$

zeigt, daß die P-Werte in jedem Punkt der Entwicklung proportional sowohl zu dem bereits erreichten Anteil wie auch zu dem noch nicht ausgeschöpften Potential (1 - P) zunehmen, das Steigungsmaximum also bei $P = (1 - P) = 0.5$ erreicht wird. Diese Eigenschaft macht die logistische Funktion auch zu einem interessanten Instrument für Untersuchungen zur Diffusion von Innovationen oder Nachrichten. Je größer die Zahl derer, die eine Nachricht schon empfangen haben, um so höher zunächst die Verbreitungsgeschwindigkeit. Ist aber erst einmal die Hälfte der potentiellen Adressaten erreicht, wird die Zuwachsrate geringer, weil die Zahl der Eingeweihten die Zahl der Nicht-Eingeweihten übersteigt.

Bevor wir uns im nächsten Abschnitt 12.2.4 mit der Frage beschäftigen, wie man die Parameter der logistischen Regressionsgleichung (12-19) bzw. (12-17) schätzen kann, wollen wir hier das Ergebnis unseres Analyse-Beispiels vorstellen, um die bisherige Diskussion zu veranschaulichen. Dabei rechnen wir nicht mit Prozentwerten der SPD-Stimmenanteile und der industriell Beschäftigten, sondern mit den relativen Häufigkeiten. Die folgende Abbildung 12.13 zeigt die logistische Regressionslinie $P = e^z/(1 + e^z)$, wobei

(12-21)

$$P = E(Y \mid X=x)$$

$$z = a + bx \;,\; a = -2.48,\; b = 3.58;\; a = \hat{\alpha},\; b = \hat{\beta}$$

Y := SPD-Stimmenanteil

x := Beschäftigtenanteil in der Industrie

In das Schaubild ist außerdem die Regressionsgerade des linearen Modells (vergl. Abb. 10.2 und 10.5) mit dem Ordinatenabschnitt a = -0.014 (-1.4 %) und b = 0.71 eingezeichnet. Die Ergebnisse der logistischen und der linearen Regression, also die jeweils ermittelten Erwartungswerte $E(Y \mid x_i)$, unterscheiden sich im Bereich eines Stimmenanteils zwischen .20 und .50 nur geringfügig beim Vergleich der beiden Modelle. Es wird deutlich, daß man mit dem logistischen Modell auch eine Beziehung darstellen kann, die innerhalb des realisierten Wertebereichs der Y- und X-Variablen nahezu linear verläuft. Da man Linearität bei qualitativen abhängigen Variablen aber nicht voraussetzen kann, ist das logistische Modell in den meisten Fällen theoretisch angemessener als das lineare Wahrscheinlichkeitsmodell. Es läßt sich ohne weiteres auf mehrere Regressorvariablen ausdehen. Auch polytome abhängige Variablen können mit ihm analysiert werden (siehe Aldrich/Nelson 1984, S. 44 ff.; Hanushek/Jackson 1977, S. 210 ff.).

Der Parameter b = 3.58 der logistischen Regression ist nicht unmittelbar als Steigungskoeffizient interpretierbar, sein Vorzeichen indiziert jedoch die positive oder negative Richtung der Variablenbeziehung (im Falle von polytomen abhängigen Variablen ist dies nicht unbedingt gewährleistet). Die punktuelle Steigung b* der logistischen Kurve variiert mit dem Niveau der X- bzw. P-Werte. Man erhält sie, indem man die Gleichung (12-17) nach X ableitet:

$$(12\text{-}22) \qquad \frac{\partial P}{\partial x} = P(1-P) \cdot b = b^*$$

Dies bestätigt ein früheres Ergebnis, wonach die maximale Steigung bei $P = (1 - P) = 0.5$ erreicht wird. In unserem Beispiel ist sie mit $b^* = 0.25 \cdot 3.58 = 0.89$ gegeben, und zwar in dem Punkt mit den Koordinaten $P_i = 0.5$ und $z_i = 0$ bzw. $x_i = (-a)/b = 2.48/3.58 = 0.69$. Bei $x_i = 0.4$, um einen anderen Punkt herauszugreifen, erhält man nach Gleichung (12-17) den Wert

$$(12\text{-}23) \qquad P_i = \frac{e^{-2.48+3.58(0.4)}}{1+e^{-2.48+3.58(0.4)}} = 0.26$$

Folglich beträgt dort die Steigung $b^* = 0.26 \cdot 0.74 \cdot 3.58 = 0.69$. Der a-Parameter verschiebt die Kurve entlang der X-Achse; der b-Parameter bestimmt, wie steil sie steht.

Die logistische S-Kurve verläuft also symmetrisch um $z_i = 0$ bzw. $x_i = (-a)/b = 0.69$. Dieser Wert ist aber im allgemeinen nicht identisch mit dem beobachteten arithmetischen Mittel $\bar{x}$. Die Kurve nähert sich zwar den Grenzwerten 0 und 1 mit $X \to \pm \infty$; innerhalb des tatsächlich beobachteten Bereichs der X-Werte wird aber nur ein Kurvenabschnitt realisiert, der in sich nicht symmetrisch verlaufen muß.

Leider läßt sich die Güte der Anpassung des logistischen Modells an die beobachteten Daten nicht mit der Anpassungsgüte des entsprechenden linearen Modells vergleichen. Der Determinationskoeffizient R^2 ist für das logistische Modell als Maß für die Stärke des Zusammenhangs problematisch, da die geschätzten Residuen mit den geschätzten Vorhersagewerten korrelieren können und somit die übliche Varianzzerlegung nicht möglich ist. Es gibt aber Vorschläge, die Anpassungsgüte über »Pseudo-Determinationskoeffizienten« auszudrücken. Darüber informieren Aldrich und Nelson (1984, S. 57 f.), Hensher/Johnson (1981, S. 48 ff.; Kühnel et al. (1989, S. 57 ff.).

12.2.4 Schätzung der logistischen Regressionsparameter mit Hilfe der Maximum-Liklihood-Methode

Bei der linearen Regression haben wir die Parameter nach dem Prinzip der Kleinsten Quadrate (OLS) ermittelt: Die Regressionsparameter wurden so gewählt, daß die quadrierten Abweichungen zwischen den beobachteten und den mit Hilfe des Modells geschätzten Y-Werten in der Summe ein Minimum bildeten. Diese Aufgabe wurde mit Hilfe der Differentialrechnung gelöst. Die Bedingungen, unter denen die OLS-Methode zu optimalen Schätzungen führt, sind bei Anteilswerten oder dichotomen Variablen im allgemeinen nicht gegeben. Deshalb werden die α- und β-Parameter des logistischen Modells (12-17) üblicherweise mit der sog. Maximum-Liklihood-Methode (MLM) geschätzt. Das Prinzip der MLM besteht darin, Populations- bzw. Modellparameter so zu schätzen, daß die bei einer Stichprobenziehung beobachteten Daten (»Ereignisse«) mit maximaler Wahrscheinlichkeit bzw. Wahrscheinlichkeitsdichte aus dieser (hypothetischen oder empirischen) Population hervorgegangen sein könnten. Unter bestimmten Voraussetzungen (die man bei der linearen Regression in der Regel als gegeben betrachtet), führen ML- und OLS-Schätzmethoden zu den gleichen Ergebnissen (siehe Abschn. 12.2.6).

Wir wollen die MLM zunächst anhand des vertrauten Beispiels einer 1/0-kodierten binären Variablen (»Bernoulli-Variable«) erläutern. Nehmen wir an, in einer Population von Wahlberechtigten befände sich ein unbekannter Anteil π von SPD-Anhängern. Die Variable Y, »Parteipräferenz«, sei binär kodiert: mit »1« für die SPD-Präferenz, mit »0« für alle anderen Präferenzen. Nehmen wir weiter an, in einer Zufallsstichprobe

von $n = 10$ Wahlberechtigten erhielten wir folgende Sequenz von Parteipräferenzen:

0 0 1 0 1 0 0 0 1 0

Der beobachtete Anteil von SPD-Änhängern in dieser Stichprobe beträgt also $p = 0{,}3$. Es liegt nahe, den unbekannten Parameter π mit $p = \hat{\pi} = 0{,}3$ zu schätzen. Wir wollen uns aber nicht auf unsere Intuition verlassen, sondern für die Wahl dieses Schätzers eine statistische Begründung gemäß dem ML-Prinzip liefern. Unsere Aufgabe besteht also darin, denjenigen Schätzer $\hat{\pi}$ zu finden, der die Wahrscheinlichkeit für die beobachtete Sequenz von Ereignissen (Parteipräferenzen) maximiert.

Zunächst probieren wir »aufs Geratewohl« einige Schätzer aus und ermitteln die daraus folgenden Wahrscheinlichkeiten, P(Pr), für die in der Stichprobe realisierten Parteipräferenzen. Wie diese theoretischen Wahrscheinlichkeitswerte mit Hilfe des Multiplikationstheorems errechnet werden, haben wir in Kap. 6.3 und 7.2 erläutert. Für die obige Ereignissequenz erhalten wir folgende Gleichung

(12-24)

$$P(Pr) = (1-p)(1-p)(p)(1-p)(p)(1-p)(1-p)(1-p)(p)(1-p)$$

$$= p^3(1-p)^7$$

Für die Schätzgröße $p = \hat{\pi}$ setzen wir nun unterschiedliche Werte ein:

p	$p^3(1-p)^7$	P(Pr)
0,1	$(0{,}1)^3(0{,}9)^7$	= 0,000478
0,2	$(0{,}2)^3(0{,}8)^7$	= 0,001678
0,3	$(0{,}3)^3(0{,}7)^7$	= 0,002357
0,4	$(0{,}4)^3(0{,}6)^7$	= 0,001792
0,5	$(0{,}5)^{10}$	= 0,000977
0,6	$(0{,}6)^3(0{,}4)^7$	= 0,000354
0,7	$(0{,}7)^3(0{,}3)^7$	= 0,000075
0,8	$(0{,}8)^3(0{,}2)^7$	= 0,000007
0,9	$(0{,}9)^3(0{,}1)^7$	= 0,000001

Eine maximale Wahrscheinlichkeit für die obige Sequenz von Parteipräferenzen wird für den Schätzer $p = 0.3$ mit $P(Pr) = 0{,}002357$ errechnet. Damit bestätigt sich unsere intuitive Wahl des Schätzers $\hat{\pi} = p$ für den Populationsparameter π. Die Funktion $P(Pr) = p^3(1-p)^7 = L$ bezeichnet man als »Likelihoodfunktion« der Stichprobe (0010100010) oder einer (einzigen) anderen Stichprobe, in der bei zehn Ziehungen dreimal das interessierende Ereignis und siebenmal das Komplementärereignis auftritt[2]. Der deutsche Ausdruck »Mutmaßlichkeit« ist kaum gebräuchlich.

[2] Die Wahrscheinlichkeit, daß das interessierende Ereignis bei n Versu-

Mathematisch allgemein findet man den ML-Schätzer, indem man mit Hilfe der Differentialrechnung das Maximum der Likelihoodfunktion bestimmt. Für die Likelihoodfunktion unseres Beispiels mit $L = p^3(1-p)^7$ und deren Logarithmus log(L) erhalten wir folgende stetige Kurve (siehe Abb. 12.14).

Das Maximum findet man über die 1. Ableitung (da es keine weiteren Maxima oder Minima gibt, ist die 2. Ableitung entbehrlich). In der Regel leitet man nicht die Likelihood-Funktion, sondern die Log-Likelihoodfunktion ab. Da die log(L)-Funktion eine monotone Transformation von L ist (siehe Abb. 12.14), haben beide ihr Maximum an der gleichen p-Stelle. In unserem Beispiel ist

$$\log(L) = 3\log(p) + 7\log(1-p) \tag{12-25}$$

Als 1. Ableitung ergibt sich daraus

$$\frac{\partial \log(L)}{\partial p} = \frac{3}{p} \cdot \frac{7}{1-p} \tag{12-26}$$

Setzt man diese Ableitungsfunktion gleich Null, erhält man

$$\begin{aligned} 3/p &= 7/(1-p) \\ p &= 0.3 \end{aligned} \tag{12-27}$$

Die Likelihoodfunktion ist formal identisch mit der Wahrscheinlichkeitsfunktion bzw. (bei stetiger Zufallsvariable) der Wahrscheinlichkeitsdichtefunktion. Im Falle der Wahrscheinlichkeits(dichte-)funktion betrachten wir den oder die interessierenden Parameter (hier: π) als gegeben und mögliche Stichprobenbeobachtungen als Variablen. Im Falle einer Likelihoodfunktion kehren wir diese Perspektive um: Die Stichprobenbeobachtungen sind gegeben, und wir betrachten den oder die Parameter als Variable, für die wir diejenigen Werte (Ausprägungen) suchen, die bei den vorliegenden Beobachtungen die Wahrscheinlichkeit, also die Likelihoodfunktion maximieren.

Wenn wir von unserem fiktiven Beispiel mit $n = 10$ und $p = 0.3$ abstrahieren, können wir die Likelihoodfunktion einer 1/0-kodierten Variablen Y wie folgt schreiben:

---

chen a-mal auftritt, ist natürlich $P = \pi^a(1-\pi)^{n-a}$. Bei der Lösung des Schätzproblems fragen wir aber nur nach der Wahrscheinlichkeit, daß bei gegebenem π die in der **einen** Stichprobe realisierten Beobachtungen in einer bestimmten Sequenz auftreten.

$$(12\text{-}28) \qquad L = \prod_{i=1}^{n} p^{y(i)} \cdot (1-p)^{1-y(i)} \qquad , \; p = \hat{\pi}$$

Das Symbol Π bedeutet, daß man die Multiplikation $p^{y(i)}(1-p)^{1-y(i)}$ n-mal durchführen soll. Da y_i nur die Werte 0 oder 1 annimmt, wird das Produkt nach dem Multiplikationssymbol für jedes i entweder gleich p (wenn $y_i = 1$) oder (1-p) (wenn $y_i = 0$). Bei $n = 10$ und $p = 0.3$ wird aus (12-28) somit (12-24).

Soll mit der dichotomen abhängigen Variablen Y ein logistisches Regressionsmodell geschätzt werden, müssen für die Anteilswerte $p = \hat{\pi}$ die entsprechenden logistischen Funktionsausdrücke in (12-28) eingesetzt werden (vergl. (12-17)):

$$(12\text{-}29) \qquad L = \prod_{i=1}^{n} \left(\frac{e^{a+bx(i)}}{1+e^{a+bx(i)}} \right)^{y_i} \cdot \left(1 - \frac{e^{a+bx(i)}}{1+e^{a+bx(i)}} \right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{a+bx(i)}}{1+e^{a+bx(i)}} \right)^{y_i} \cdot \left(\frac{1}{1+e^{a+bx(i)}} \right)^{1-y_i}$$

Daraus ergibt sich die folgende logarithmierte Likelihood-Funktion

(12-30)

$$\log(L) = \sum_{i=1}^{n} \left[y_i \cdot \log\left(\frac{e^{a+bx(i)}}{1+e^{a+bx(i)}} \right) + (1-y_i) \log\left(\frac{1}{1+e^{a+bx(i)}} \right) \right]$$

Die Parameter a und b sind nun so zu wählen, daß log(L) ein Maximum wird. Wir ersparen es uns, die partiellen Ableitungen nach a und b hier nachzuvollziehen. Es ist auch so erkennbar, daß sie nicht linear sind. Statt einer analytischen wird somit eine numerisch-iterative Lösung notwendig. (Ein Verfahren hierzu ist in Linder/Berchtold 1976, S. 57 ff.) skizziert.) Normalerweise werden dabei auch die Varianzen und Kovarianzen der Regressionsparameter ermittelt. Das in SPSS^x/CNLR implementierte Verfahren berechnet die Standardfehler jedoch nach der sog. Bootstrapping-Methode, die wir hier nicht näher erläutern wollen. Im Prinzip geht man

dabei so vor, daß aus dem gegebenen Datensatz wiederholt (z. B. dreißigmal) Stichproben (mit »Zurücklegen«) gezogen und die Parameter für jede Stichprobe gesondert berechnet werden. Auf diese Weise erhält man eine empirische Verteilung des jeweiligen Regressionsparameters, aus der man die Standardabweichung bestimmen kann.

Bevor wir das SPSS^x-Programm hierzu erläutern, müssen wir noch die Likelihoodfunktion für Prozent- bzw. Anteilswerte kurz erörtern, da wir es bei den Wahlkreisdaten in unserem Beispiel nicht mit binär kodierten abhängigen Variablen, sondern mit Stimmenanteilen zu tun haben. Die Prozentdaten (p_j) stellen gewichtete Summen der pro Wahlkreis anfallenden 1/0-kodierten Individualwerte y_{ij} dar, wobei sich das Gewicht aus der Zahl n_j der Wahlberechtigten im Wahlkreis j ergibt:

$$(12\text{-}31) \qquad p_j = (100/n_j)(y_{1j} + y_{2j} + \ldots + y_{nj}) = (100/n_j) \cdot S_j$$

(Bei reinen Anteilswerten wird der Faktor 100 zum Faktor 1.) Die Prozent- oder Anteilswerte sind also wie die Summen (S_j) der binär kodierten Variablen binomial verteilt gemäß Gleichung (7-2):

$$(12\text{-}32) \qquad \binom{n_j}{S_j} \hat{\pi}_j^{s(j)} \cdot (1-\hat{\pi})^{n(j)-s(j)} = f(s_j)$$

Der einzige formale Unterschied zur Gleichung (12-24) besteht in dem Faktor $\binom{n_j}{S_j}$. Wie wir in Kap. 7.2 gesehen haben, ergibt er sich aus der Tatsache, daß die Summe S_j (und damit die Anteilswerte) aus verschiedenen Kombinationen von Individualpräferenzen hervorgegangen sein kann, wobei jede einzelne die Wahrscheinlichkeit $P_{ij} = \pi_j^{S(j)}(1 - \pi_j)^{n(j) - S(j)}$ hat.

Die Likelihoodfunktion erhalten wir wiederum, indem wir nicht nur für eine einzelne Summe S_j bzw. den Anteilswert p_j, sondern für die Gesamtheit der beobachteten Daten über alle n Wahlkreise $j = 1, 2, \ldots, n = 397$ die gemeinsame Verteilung angeben:

$$(12\text{-}33) \qquad L = \prod_{j=1}^{n} \binom{n_j}{S_j} \pi^{s(j)}(1-\pi_j)^{n(j)-s(j)}$$

Der (natürliche) Logarithmus daraus ist

$$(12\text{-}34) \qquad \ln(L) = \sum_{j=1}^{n} \ln\binom{n_j}{S_j} + S_j \ln\pi_j + \sum_{i=1}^{n} (n_j - S_j)\ln(1-\pi_j)$$

Die Konstanten $\Sigma \ln \binom{n_j}{S_j}$ werden in den Ableitungen nach π gleich Null, so daß sie vernachlässigt werden können. Außerdem kann man die Faktoren S_j und $(n_j - S_j)$ durch n_j dividieren, so daß man direkt die Anteilswerte $p_j = S_j/n_j$ bzw. $(1-p_j) = (n_n - S_j)/n_j$ einsetzen kann. Außerdem substituieren wir für die Wahrscheinlichkeiten $\hat{\pi}_j$ die logistische Funktion (12-17), so daß wir wiederum eine logarithmische Likelihoodfunktion erhalten, die Gleichung (12-30) entspricht[3]; es werden lediglich die 1/0-kodierten Y-Werte durch die Anteilswerte p_i ersetzt:

(12-34a)

$$\ln(L) = \sum_{i=1}^{n} p_j \ln \frac{e^{a+bx(j)}}{1+e^{a+bx(j)}} + \Sigma(1-p_j)\ln\left(1 - \frac{e^{a+bx(j)}}{1+e^{a+bx(j)}}\right)$$

Die zu maximierende Log-Funktion muß (in modifizierter Form) in das SPSS^x-Programm für die logistische Regression eingegeben werden. Dieses Programm wollen wir nun anhand unseres Beispiels erläutern.

12.2.5 Durchführung der logistischen Regression mit dem EDV-Programmpaket SPSS^x (Version 3.1)

Die logistische Regression ist eine bestimmte Variante der nicht-linearen Regression, für die SPSS^x Schätzalgorithmen sowohl nach der MLM als auch nach der Kleinstquadratmethode anbietet. Wir erläutern hier aber nur die Ausführung nach der MLM. Sie ist sowohl bei dichotomen (im Prinzip, aber leider nicht in SPSS^x, auch bei polytomen) abhängigen Variablen wie auch beim Rechnen mit Anteilswerten anwendbar.

Nach einem mit MODEL PROGRAM eingeleiteten Kommando (siehe unten) muß zunächst diejenige nicht-lineare Funktion angegeben werden, deren Parameter geschätzt werden sollen. In unserem Beispiel also die logistische Funktion (vergl. (12-17)):

[3] Es fällt auf, daß in der Logit-Gleichung kein (stochastischer) Fehlerterm vorgesehen ist. Es wird also vorausgesetzt, daß bei gegebenem X-Wert die Wahrscheinlichkeiten P(Y = 1) für alle Fälle gleich sind (siehe Hanushek/Jackson 1977, S. 203).

```
COMPUTE PRED=EXP(A+B*INDANT)/(1+EXP(A+
                B*INDANT)
```

PRED ist der Default-Name für das zu schätzende Modell, EXP steht für das Exponieren zur Basis der Eulerschen Zahl e. Die Industrialisierungsvariable INDANT wurde (aus Gründen, die gleich erläutert werden) nicht als Prozentgröße, sondern als Anteilswert eingesetzt.

Als nächstes ist die sog. »Loss«-Funktion (die Minimierungsfunktion) anzugeben. Ihrem Inhalt nach handelt es sich dabei um die logarithmierte Likelihoodfunktion, wie wir sie oben erörtert haben. Sie wird nun allerdings formal modifiziert: Es wird nicht die Log-Likelihood maximiert, sondern ihr negativer Ausdruck minimiert. Außerdem muß man bedenken, in welchen Wertebereich eine Potenzzahl e^{a+bx} mit $e \approx 2{,}718$ vorstoßen kann; e^{80} z. B. wird die Speicherkapazität eines Rechners überschreiten. Es ist deshalb günstig, die Variablen monoton so zu transformieren, daß kleinere Skalenwerte zustande kommen. (Der Punkt, bei dem die Loss-Funktion ihr Minimum erreicht, wird dadurch in der Horizontalen nicht verschoben.) Bei Prozentangaben bietet sich eine Division durch 100 an, so daß man mit Anteilswerten $0 \leqq p \leqq 1$ rechnen kann. Wir transformieren also in dem Programmabschnitt, der die Datenmodifikationsbefehle enthält (vor den Prozedurbefehlen) mittels entsprechender COMPUTE-Statements die Prozentangaben sowohl der SPD- als auch der INDUSTRY-Variablen in Anteilswerte und schreiben die Loss-Funktion wie folgt (vergl. (12-34a)):

```
COMPUTE LOSS=( - SPDANT*LN(PRED)) - ((1 - SPDANT)*
                LN(1 - PRED))
```

Abhängige und unabhängige Variablen werden im CNLR-Subkommando spezifiziert, das noch weitere Befehle aufnehmen kann, z. B.:

```
CNLR SPDANT WITH INDANT/PRED=PRED/LOSS=LOSS/
  BOOTSTRAPS/SAVE=PRED,RESID
```

Nach dem Schlüsselwort CNLR dürfen nur Variablen gelistet werden, die als Argumente in den vorangegangenen Funktionsgleichungen PRED und LOSS aufgetreten und im aktiven File vorhanden sind, also keine mit diesen Gleichungen kreierten temporären Variablen. Das PRED-Subkommando (innerhalb des CLNR-Befehls) ist entbehrlich, wenn zuvor in der Modellspezifikation (COMPUTE PRED) die Default-Variable PRED und nicht irgendeine andere Bezeichnung benutzt wurde. Die Voreinstellung für die LOSS-Funktion in CNLR ist die Minimierung der Fehlerquadratsumme. Mit LOSS=LOSS wird statt dessen die mit COMPUTE LOSS zuvor spezifizierte Funktion minimiert.

Wie bereits angedeutet, erhält man die Standardfehler der Regressionsparameter mit Hilfe des BOOTSTRAP-Kommandos. In unserem Beispiel

werden sie mit $s_a = 0.146$ (a = - 2,48) und $s_b = 0.343$ (b = 3.58) errechnet. Sie sind in gleicher Weise zu interpretieren wie bei der linearen Regression (siehe Abschn. 10.3). Konfidenzintervalle für die Regressionsparameter werden ebenfalls in der dort beschriebenen Weise konstruiert. Eine Teststatistik für die Anpassungsgüte des Modells liefert SPSS^x/CNLR nicht direkt; sie läßt sich jedoch aus anderen Angaben errechnen (zur Beschreibung einiger dieser Testverfahren siehe Aldrich/Nelson 1984, S. 55 f.; Hensher/Johnson 1981, S. 48 ff.; Kühnel et al. 1989, S. 61ff.)

Mit dem SAVE-Subkommando können u. a. die von dem Modell prognostizierten Werte sowie die Residuen gespeichert und in nachfolgenden Prozeduren (z. B. im PLOT-Kommando) weiter verarbeitet werden. Der CNLR-Befehl kann weitere optionale Subkommandos aufnehmen, die im Manual nachzulesen sind.

Damit der Schätzalgorithmus überhaupt in Gang gesetzt werden kann, müssen für die iterativ zu schätzenden Parameter Startwerte angegeben werden. Das geschieht mit dem ersten Befehl, der das Programm zur nichtlinearen Regression aufruft:

```
MODEL PROGRAM A= - 3.4, B=5.17
```

Diese Anfangswerte kann man auf unterschiedliche Weise ermitteln (siehe SPSS^x-Manual, S. 679, 708 f.). Sie sollten in den Zehnerpotenzen den endgültigen Schätzwerten entsprechen. Die in der Regel einfachste Methode ist es, den Fehlerterm der (logistischen) Regressionsgleichung zu ignorieren und die übrige Modellgleichung, falls möglich, zu linearisieren.

Wir haben schon oben dargelegt, daß die Logitgröße z = ln [SPDANT/(1 - SPDANT)] linear mit der Industrialisierungsvariable verknüpft ist, die wir ebenfalls als Anteilsvariable, INDANT, definiert haben. Unter der Voraussetzung, daß die Anteilswerte null und eins nicht vorkommen, können wir die Startwerte für die logistische Regression über folgende lineare Regression ermitteln:

```
COMPUTE Z=LN((SPDANT/(1 - SPDANT))
REGRESSION VARIABLES=Z, INDANT/DEP=Z/ENTER
```

In unserem Beispiel erhalten wir die Startwerte: $a^* = - 3.4$ (Ordinatenabschnitt) und $b^* = 5.17$ (Steigungskoeffizient).

Falls Anteilsgrößen von 0 oder 1 vorkommen, muß ein kleiner Umweg beschritten werden, der im SPSS^x-Manual (3rd edition) beschrieben wird.

Die iterative Schätzung der logistischen Regressionskoeffizienten kann durch verschiedene Programmparameter innerhalb des CNLR-Befehls gesteuert werden (siehe dazu ebenfalls das Manual). Wir haben in unserer Beispiel-Rechnung die Default-Kriterien benutzt.

Hier noch einmal der gesamte Programmaufbau zur logistischen Regression im Überblick:

```
COMPUTE GEWICHT=(BERECH12/14366567)*395
WEIGHT BY GEWICHT
MODEL PROGRAM A=-3.4 B=5.17
COMPUTE PRED=EXP(A+B*INDANT)/(1+EXP(A+
                B*INDANT))
COMPUTE LOSS=( - SPDANT*LN(PRED)) - ((1 - SPDANT)*
                LN(1 - PRED))
CNLR SPDANT WITH INDANT/ PRED=PRED /LOSS=
     LOSS /BOOTSTRAPS /SAVE=PRED,RESID
```

Das Verfahren ist ohne zusätzliche technische Schwierigkeiten auf mehrere Regressorvariablen ausdehnbar, indem der Exponent in e^{a+bx} auf $e^{a+\Sigma b_k x_k}$ erweitert wird (siehe hierzu das Textbeispiel in Aldrich/Nelson 1984 und das dazugehörige SPSS^x-Programm im Manual, 3rd edition, S. 710 ff.). Die Befehle MODEL PROGRAM und COMPUTE PRED sind rechts vom Gleichheitszeichen entsprechend zu ergänzen, z.B.

```
MODEL PROGRAM A = - 3.0 B1=3.2 B2=2.5
COMPUTE PRED=EXP(A+ B1 * INDANT + B2 * EVANT)/
                1 + EXP(A + B1 * INDANT + B2 * EVANT)
```

Erweitert man das Modell auf mehrere Regressorvariablen, ist man an deren relativer Erklärungskraft interessiert. Dazu sind sog. Effektparameter konstruiert worden, die Kühnel et al. (1989) erläutern. Zu beachten ist, daß das logistische Regressionsmodell inhärent multiplikativ ist: die Wirkung von X_1 auf Y hängt stets davon ab, welche Werte in der anderen Regresssorvariable X_2 erreicht sind. Auch dieser Aspekt wird in Kühnel et al. (1989) erläutert.

Leider ist SPSS^x/CLNR derzeit auf Modelle mit dichotomen abhängigen Variablen (oder die daraus resultierenden Anteilswerte) begrenzt. Modelle mit polytomen abhängigen Variablen (die in **mehrere** Dummy-Variablen aufzulösen wären) sind nicht schätzbar. Außerdem fehlen wichtige Teststatistiken, insbesondere für die Anpassungsgüte, oder sie können nur über zusätzliche Berechnungen konstruiert werden.

Logistische Regressionen mit MLM-Schätzung können in SPSS^x auch noch mit einem anderen Unterprogramm, nämlich dem PROBIT-Befehl gerechnet werden. Technisch ist dies sogar einfacher zu bewerkstelligen, weil keine Lossfunktion angegeben werden muß:

```
PROBIT HSPD12 OF BERECH12 WITH INDUSTRY
     /MODEL=LOGIT /LOG=NONE
```

Die PROBIT-Prozedur (mit der man neben den Logit- auch Probit-Modelle rechnen kann) erwartet als Eingabe **gruppierte** Daten einer 1/0-kodierten abhängigen Variablen. Die Zielvariable, die angibt, wie oft

die Ausprägung Y = 1 in der »Gruppe« vorkommt, folgt unmittelbar dem PROBIT-Aufruf. In unserem Beispiel ist die Anzahl (nicht der Anteil) der SPD-Stimmen in der Variablen HSPD12 für jeden Wahlkreis registriert. Hinter dem OF folgt die Variable, die für alle »Gruppen« (in unserem Beispiel die Wahlkreise) die Zahl der Fälle angibt, für die eine Y-Ausprägung (1 oder 0) beobachtet wurde. In unserem Beispiel ist das die Anzahl der Wahlberechtigten in BERECH12. Nach dem WITH folgt die Regressorvariable, hier der Industrialisierungsgrad der Wahlkreise. Mit MODEL = LOGIT wird der logistische Funktionstyp aufgerufen; mit LOG = NONE wird verhindert, daß die Variablen vorher logarithmiert werden (was das Programm sonst als Voreinstellung ausführt).

Das PROBIT-Programm liefert die Regressionskoeffizienten a und b nach einer Transformation der Daten, die aus rechentechnischen Gründen programmintern vorgenommen wird. Bei der Interpretation der Parameter, wie wir sie oben erläutert haben, muß diese Transformation rückgängig gemacht werden. Der errechnete b'-Koeffizient (in unserem Beispiel b' = 1.79) muß mit 2 multipliziert werden (b = 2 · 1.79 = 3.58), der a-Koeffizient (in unserem Beispiel a' = 3.76) muß zuerst mit 2 multipliziert, anschließend um den Betrag 10 vermindert werden: a = (3.76 · 2) - 10 = - 2.18. Diese Koeffizienten stimmen mit denen überein, die wir über das CLNR-Programm ermittelt hatten. Auch die ausgedruckten Standardfehler müssen mit 2 multipliziert werden. Das PROBIT-Programm liefert einen Chi-Quadrat-Test für die Anpassungsgüte, der aber nur als Vergleichsgröße sinnvoll ist, wenn man mehrere Modelle (mit unterschiedlichen Sätzen von Regressorvariablen) schätzt. Er ist lediglich für die Analyse mit gruppierten Daten, nicht den originären Individualdaten gedacht. Bei der Analyse von Individualdaten können weitere technische Probleme auftreten. Außerdem wird die Likelihoodfunktion nicht ausgedruckt. Wegen dieser Programmeinschränkungen wurde hier das allgemeinere SPSS^x/CLNR-Programm eingeführt, das die logistische Regression als eine neben anderen Formen nicht-linearer Regression behandelt.

12.2.6 Exkurs: Die ML-Schätzer für das lineare Regressionsmodell

In diesem Exkurs wollen wir zeigen, wie man die MLM auch zur Schätzung der Regressionsparameter des **linearen** Modells anwenden kann. Auf diese Weise soll das Verständnis der MLM weiter vertieft werden, da viele der hier nicht behandelten komplexen statistischen Verfahren sich dieser Schätzmethode bedienen.

Wir gehen davon aus, daß die in der Stichprobe beobachteten Werte (y_1, y_2, ..., y_n) der kontinuierlichen Zufallsvariablen Y normalverteilt um ihre bedingten Erwartungswerte $E(Y \mid x_i) = \alpha + \beta x_i$ streuen. Für jede einzelne Beobachtung gilt somit die folgende Wahrscheinlichkeitsdichtefunktion:

(12-35)

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - E(y_i))^2}{2\sigma^2}} \qquad , \pi \approx 3{,}14159$$

$$= (2 \cdot \pi \cdot \sigma^2)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(y_i - E(y))/\sigma]^2}$$

$$= (2 \cdot \pi \cdot \sigma^2)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}[(y_i - \alpha - \beta x_i)/\sigma]^2}$$

Die gemeinsame Dichtefunktion für alle n Beobachtungen, geschrieben als Likelihoodfunktion, ergibt sich aus dem Multiplikationstheorem der Wahrscheinlichkeitsrechnung mit

(12-36) $\quad L = f(y_1) \cdot f(y_2) \cdot \ldots \cdot f(y_n)$

logarithmiert zu

(12-37)

$$\log(L) = \sum_{i=1}^{n} \log[f(y_i)]$$

$$\log[f(y_i)] = -\tfrac{1}{2}\log(2\pi \cdot \sigma^2) - \tfrac{1}{2}[(y_i - \alpha - \beta x_i)/\sigma]^2$$

$$-\log(2\pi\sigma^2) = -\log(2\pi) - \log(\sigma^2)$$

$$\log(L) = -\frac{n}{2} \cdot \log(2\pi) - \frac{n}{2} \cdot \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Für die drei unbekannten Parameter α, β und σ^2 erhält man drei Lösungsgleichungen, wenn man log(L) schrittweise nach α, β und σ^2 partiell ableitet. Im Hinblick auf α und β führen diese Ableitungen und deren Gleichsetzung mit Null zu den Normalgleichungen (10-7) der Kleinstquadratmethode. Hinsichtlich σ^2 erhält man den Schätzer $\hat{\sigma}^2 = 1/n(\Sigma e_i^2)$, der, wie wir sahen, nicht erwartungstreu ist. (Erwartungstreu ist der Schätzer $\hat{\sigma}^2 = 1/df(\Sigma e_i^2)$.) Parameter, die mit der MLM geschätzt werden, haben im übrigen die sehr wünschenswerten Eigenschaften asymptotisch effizient (siehe oben, Kap. 8.3) und asymptotisch normalverteilt zu sein.

Abb. 12.1: Beispiel einer nicht-linearen Beziehung

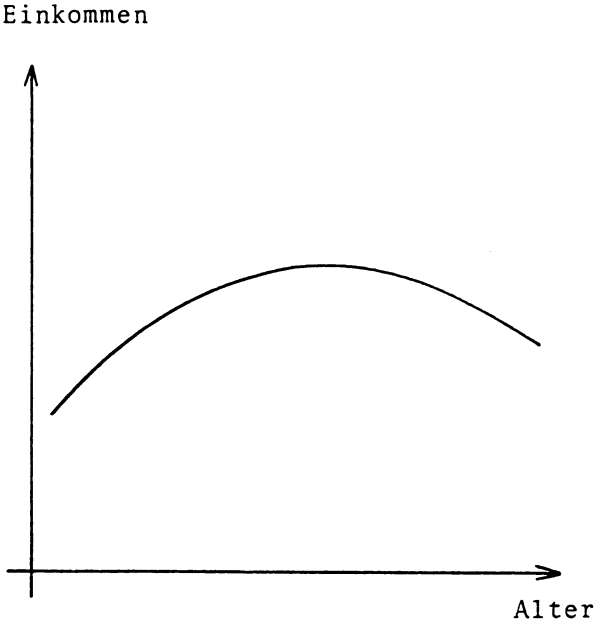
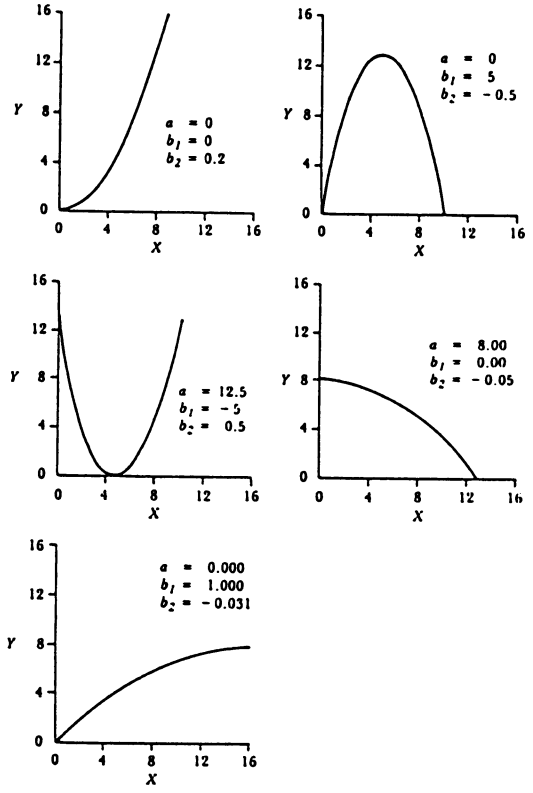
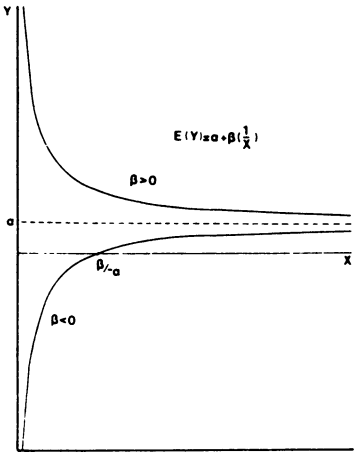


Abb. 12.2: Unterschiedliche Beziehungsformen, die mit einem Polynom 2. Grades darstellbar



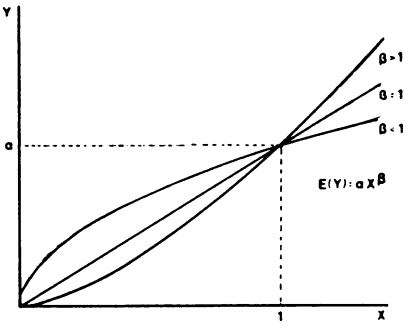
Quelle: Stolzenberg 1979, S. 465

Abb. 12.3: Hyperbolische Beziehung zwischen abhängiger u. unabhängiger Variable



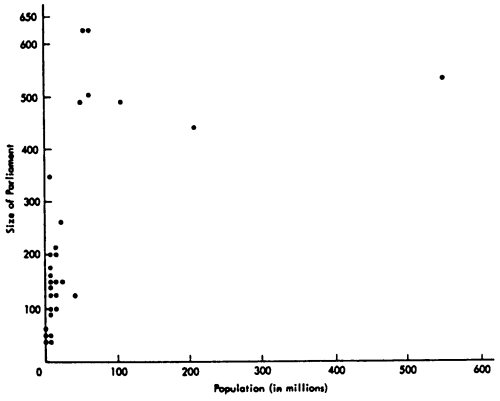
Quelle: Berry/Feldman 1985, S. 61

Abb. 12.4: Exponentialmodelle



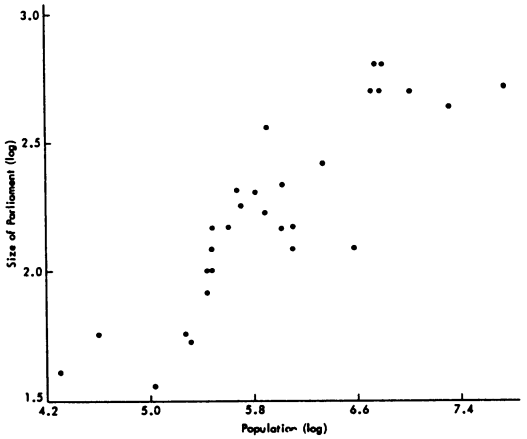
Quelle: Berry/Feldman 1985, S. 61

Abb. 12.5: Beziehung zwischen Parlamentsgröße und Bevölkerungsgröße in 29 Ländern



Quelle: Tufte 1974, S. 113

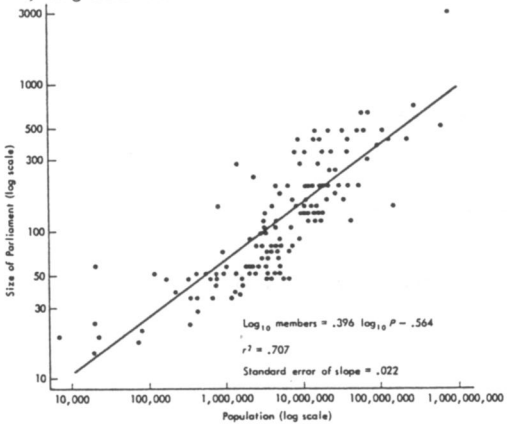
Abb. 12.6: Beziehung zwischen Parlaments- und Bevölkerungsgröße, beide Variablen logarithmiert



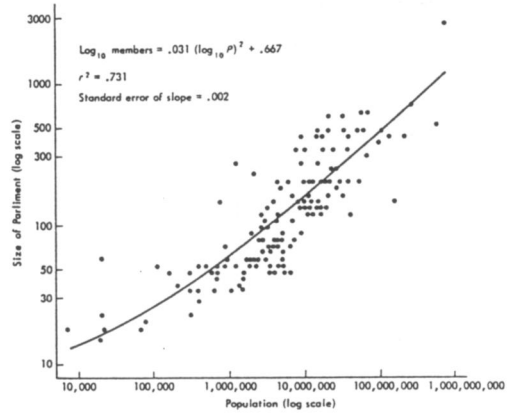
Quelle: Tufte 1974, S. 114

Abb. 12.7: Beziehung zwischen Parlaments- und Bevölkerungsgröße

a) log-lineares Modell



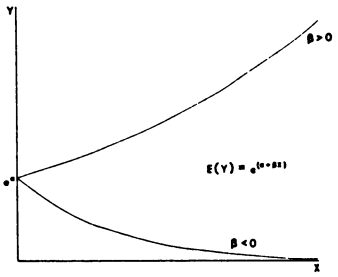
b) unabhängige Variable quadriert



Quelle: Tufte 1974, S. 116 f.

Abb. 12.8: $LogY = a + bX$

$E(Y) = e^{a+bx}$



Quelle: Berry/Feldman 1985, S. 61

Abb. 12.9: Fiktiver Zusammenhang zwischen industrieller Beschäftigung und SPD-Votum auf der Individualebene

Für SPD?	In Industrie beschäftigt? nein	ja
nein	90 %	40 %
ja	10 %	60 %
	100 %	100 %

Abb. 12.10: Muster einer S-förmigen Beziehung

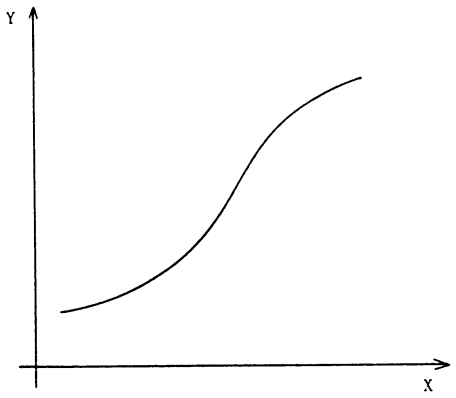
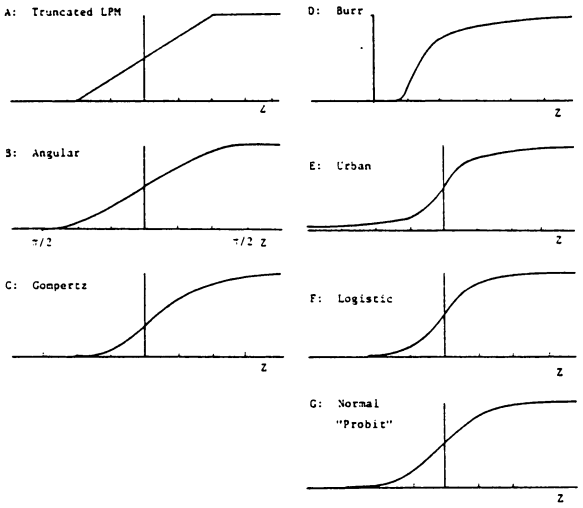
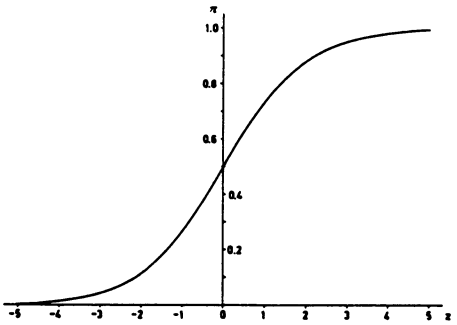


Abb. 12.11: Einige Formen nicht-linearer Regressionsmodelle



Quelle: Aldrich/Nelson 1984, S. 33

Abb. 12.12: Verlauf der Funktion $\pi = 1/(1+e^{-z})$



Quelle: Linder/Berchtold 1976, S. 29

Abb. 12.13: Logistische Regression: Prognostizierte SPD-Stimmenanteile in Abhängigkeit vom Industrialisierungsgrad

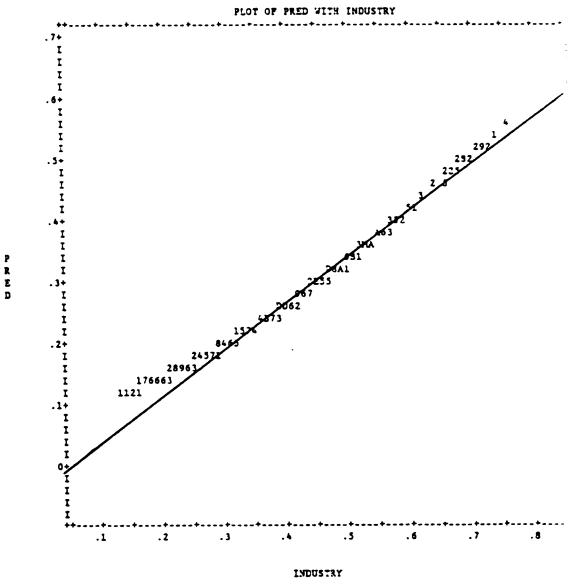
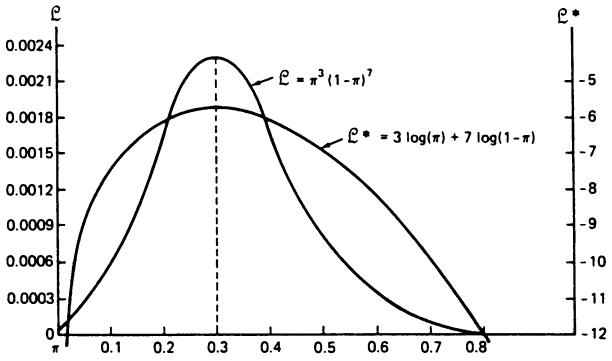


Abb. 12.14: Likelihood (L) und Log-Likelihood (L*) Funktionen



Quelle: Hanushek/Jackson 1977, S. 346

KAPITEL 13

Hinweise zum Problem fehlender Werte (*)

Unvollständige Quellen, fehlende Informationen sind ein Hauptproblem der historischen Sozialforschung. Im Rahmen eines Einführungstextes zur Statistik kann es nicht ausführlich behandelt werden. Aber ich möchte wenigstens mit ein paar Hinweisen auf die Problematik aufmerksam machen und dazu anregen, sich mit ihr weiter auseinanderzusetzen.

Wie in Kap. 2, Teil I erläutert, setzt eine statistische Analyse voraus, daß die Daten in Form einer rechteckigen Matrix gemäß Abb. 2.1 angeordnet werden können. In der Praxis treten dabei fast immer irgendwelche Lükken auf: Für eine mehr oder weniger große Zahl von Untersuchungseinheiten fehlen die Werte bei einer oder mehreren Variablen (»Missing Data« - MD). Gelegentlich fehlen für einige Fälle die Werte aller in einer bestimmten Analyse verwendeten Variablen, obgleich Angaben auf anderen Merkmalsdimensionen erhoben werden konnten.

Nicht selten sind die historischen Quellen jedoch so unvollständig, daß die Zielpopulation, über die man letztlich Aussagen machen will (siehe Kap. 9), nur teilweise erfaßt wird, ganze Gruppen von Untersuchungseinheiten, manchmal unbekannterweise, ausgeschlossen bleiben. Selbst das Zusammenlegen mehrerer Datenfiles aus verschiedenen Quellen (»record linkage«) reicht oft nicht aus, die Lücken zu schließen. Gegenüber solchen Totalausfällen sind auch die gelegentlich als Zaubermittel angesehenen inferenzstatistischen Methoden machtlos. Wenn es kein »Modell« gibt, das die Ausfälle erklärt, oder wenn über die mutmaßlichen Erklärungsfaktoren ebenfalls keine Daten vorliegen, entfallen die Anwendungsbedingungen der Inferenzstatistik. Es bleibt einem nichts anderes übrig, als seine Aussagen auf diejenige Population zu begrenzen, die (mit oder ohne Stichprobenziehung) tatsächlich erfaßt wurde. Allerdings können u. U. - vor allem, wenn die Varianz der abhängigen Variablen durch die Ausfälle begrenzt wird - nicht einmal solche Einschränkungen die Validität der Aussagen retten (siehe Abschn. 4.2.7, Teil I, und Berk 1983). Diese Problematik kann hier nicht vertieft werden.

Aber schon in den »einfacher« aussehenden Situationen, in denen einzelne Fälle lediglich auf einigen Merkmalsdimensionen keine Angaben enthalten, können erhebliche Schwierigkeiten entstehen. In der Regel werden solche Fälle ignoriert, aus der Analyse ausgeschlossen. Dieses »Verfahren« haben wir auch in den Demonstrationsbeispielen dieses Einführungstextes angewandt. In der Praxis setzt man meistens voraus, daß solche

Ausfälle »zufällig« zustandegekommen sind oder daß sie bei geringer Ausfallrate (Daumenregel: bis etwa 10 %) die Ergebnisse nicht nennenswert verzerren. Es liegen jedoch Beispiele vor (siehe Anderson et al. 1983, S. 479), die zeigen, daß auch niedrige Ausfallraten u. U. bedeutsame Ergebnisverzerrungen nach sich ziehen.

Solche Verzerrungen sind ausgeschlossen, wenn die »Missing Data« tatsächlich nur »rein zufällig« fehlen: »missing completely at random« - MCAR. Eine Begrenzung der Analyse auf Fälle mit vollständigen Daten stellt dann lediglich eine Reduktion des Stichprobenumfangs dar (was allerdings mit einem Verlust an Präzision verbunden ist). Zufälliges Fehlen (MCAR) bedeutet, daß das Auftreten von missing data bei einer bestimmten Variablen weder mit den Werten dieser Variablen noch mit den Werten anderer Untersuchungsvariablen korreliert (Kim/Curry 1977, S. 217). Diese Annahme wird z. B. verletzt, wenn Angaben zum Einkommen häufiger bei den Reicheren als bei den Ärmeren fehlen. Präzisere und vollständigere MCAR-Definitionen findet man bei Anderson et al. (1983, S. 416) oder bei Schnell (1986, S. 5 ff.).

Für das Testen auf Zufälligkeit der MD's sind eine Vielzahl von Verfahren vorgeschlagen worden (einen Überblick gibt Schnell 1986), die aber stets bestimmte Aspekte einer vollständigen MCAR-Definition außer acht lassen (ebd., S. 12). Allerdings lassen sich grobe Verstöße gegen die MCAR-Annahme in den allermeisten Fällen schon mit sehr einfachen Tests feststellen (Schnell 1988, S. 1).

Ein Verfahren zur Überprüfung der MCAR-Annahme besteht z. B. darin, eine oder mehrere Variablen, M_k ($k = 1,2,...,K$), mit einem relativ hohen Anteil an missing data so zu dichotomisieren, daß fehlende Werte mit 0, vorhandene Werte mit 1 rekodiert werden. Anschließend wird der Zusammenhang zwischen diesen Indikatorvariablen M_k und anderen Untersuchungsvariablen V_i geprüft, die keine fehlenden Werte aufweisen. Handelt es sich bei V_i um eine metrische Variable, wird der Mittelwert von V_i einmal für die Fälle mit $M_k = 1$ und zum anderen für die Fälle mit $M_k = 0$ errechnet. Die Differenz der beiden Mittelwerte sollte bei MCAR nahe Null liegen; sie kann mit Hilfe eines t-Tests auf ihre »Signifikanz« überprüft werden (siehe Kap. 8). Handelt es sich bei V_i um eine nominalskalierte Variable, läßt sich statt des t-Tests ein Chi-Quarat-Test durchführen. Will man mehrere Variablen simultan untersuchen, bieten sich die verschiedenen Methoden der mehrdimenisonalen Tabellenanalyse (bis hin zur Anwendung log-linearer Modelle) an.

Tests mit Hilfe der Indikatorvariablen lassen Verstöße gegen die Annahme der Zufälligkeit fehlender Werte nur insoweit erkennen, als die Indikatorvariablen mit dem Prozeß verbunden sind, der die missing data erzeugt hat.

Gelegentlich enthalten alle theoretisch relevante Variablen fehlende Werte. In diesem Fall läßt sich die Zufälligkeit ihres Musters überprüfen,

indem man sämtliche Indikatorvariablen (zunächst paarweise) miteinander korreliert; denn die missing data der einzelnen Merkmale sollten nicht miteinander korrespondieren. Wenn diese Annahme verletzt ist, kann die Korrelationsmatrix der Indikatorvariablen mit Hilfe komplexerer Verfahren (z. B. der Faktorenanalyse) weiter untersucht werden, um eventuell Hinweise auf die Systematik des Ausfallprozesses zu erhalten.

Für die Behandlung fehlender Werte bieten sich grundsätzlich drei Alternativen an:

(1) die entsprechenden Fälle werden aus der Analyse herausgenommen;
(2) man versucht, die fehlenden Werte anhand anderer Informationen zu schätzen, also eine vollständige Datenmatrix zu konstruieren;
(3) man versucht, die interessierenden Modellparameter (z. B. Korrelations- oder Regressionskoeffizienten) mit einer Korrektur für fehlende Werte zu schätzen, ohne diese selbst durch ein vorgeschaltetes Schätzverfahren zu ersetzen.

Für die Alternativen (2) und (3) sind eine Vielzahl formaler, statistischer Verfahren vorgeschlagen worden (siehe den Überblick bei Anderson et al. 1983 oder, umfassender, bei Schnell 1986). Ihre Leistungsfähigkeit, d. h. ihr Vermögen, zu unverzerrten Schätzgrößen zu führen, hängt aber wesentlich von dem tatsächlichen (oft unbekannten) Ausfallmechanismus ab; außerdem von weiteren Bedingungen, wie der Menge fehlender Werte und der Zahl der involvierten Variablen. Eine generelle Bewertung dieser Methoden, die für alle Anwendungsbedingungen zuträfe, ist nicht möglich, zumal auch noch der zeitliche und finanzielle Aufwand einer Analyse fehlender Werte zu berücksichtigen wäre. Die meisten Verfahren, die zur Zeit gehandelt werden, setzen den MCAR-Mechanismus voraus. Aber selbst wenn die Ausfälle zufällig entstanden sind, ist festzustellen: »Because of the complicated and interacting effects of the various factors influencing the relative success of the competing techniques, no one method for handling the missing data problem can be shown to be uniformly superior« (Anderson et al. 1983, S. 479). Schnell (1985, S. 66) resümiert die Ergebnisse einer Simulationsstudie, in der er die Effizienz verschiedener MD-Techniken unter wechselnden Bedingungen überprüft hat, in ähnlicher Weise: »Ein unter allen Bedingungen effizientes Schätzverfahren existiert nicht.«

Der Trend in der Methodenentwicklung zur Lösung oder besseren Bearbeitung dieses Problems zielt in die folgende Richtung (siehe Schnell 1985, S. 69 f.): Der Prozeß, der zum Ausfall von Informationen führt, muß, gestützt auf substanzwissenschaftliche Theorien sowie eine möglichst genaue Kenntnis der Quellengeschichte (siehe z. B. Shapiro et al. 1987) und der in ihr wirksam gewordenen Selektionsmechanismen explizit modelliert werden. Unter Umständen sind hierfür alternative Modelle zu

konstruieren. Auf der Basis dieser Modelle sind fehlende Werte wiederholt zu schätzen (»Multiple Imputation«). Der Vergleich der Ergebnisse wiederholter Schätzungen erlaubt Rückschlüsse auf die Verteilung der Statistiken bei gegebenen Daten und gegebenem Ausfallprozeß. Bei Vorgabe alternativer Modelle sollten so auch Aussagen über das Maß der Abhängigkeit der Ergebnisse von unterschiedlichen Ausfallmechanismen möglich sein. Letztlich werden also nicht so sehr eindeutige Ergebnisse, sondern eher Einsichten über die Spannweite möglicher Ergebnisse erwartet (ebd., S. 70). Allerdings »(muß) zusammenfassend betont werden, daß die Durchführung einer MI (Multiplen Imputation, H. T.) z. Z. mit einer so großen Zahl praktischer Probleme behaftet ist, daß die Anwendung im Rahmen einer Standardanalyse weitgehend ausscheidet« (ebd.)

Bei der Behandlung fehlender Werte bewegen wir uns also gegenwärtig noch auf ungesichertem Terrain. In der Forschungspraxis bieten sich jedoch gelegentlich ad-hoc Methoden an, die das Problem der fehlenden Werte zumindest mindern. So können z.B. fehlende Angaben zum formalen Schulabschluß im 19. Jahrhundert zumindest teilweise über Angaben zum erlernten Beruf »geschätzt« werden, da in dieser Zeit formale Schulbildung und Berufswahl hoch miteinander korrelierten.

Die meisten multivariaten Analysen haben ihren Ausgangspunkt in Korrelations- und Kovarianzmatrizen, in denen die bivariaten Korrelationen (Kovarianzen) aller beteiligten Variablen enthalten sind. Bei der Erstellung solcher Korrelationsmatrizen (z. B. im Rahmen einer multiplen Regression oder einer Faktorenanalyse) mit fehlenden Variablenwerten wird in der Regel eine von zwei Alternativen gewählt (in der Terminologie des Programmpakets SPSS): »listwise« oder »pairwise deletion« der betreffenden Fälle. »Listwise« bedeutet, daß sämtliche Fälle aus der Untersuchung ausgeschlossen werden, die bei mindestens einer der Untersuchungsvariablen einen fehlenden Wert aufweisen. Erfolgt der Ausschluß hingegen »paarweise«, werden zur Berechnung der (bivariaten) Kovarianzen lediglich diejenigen Fälle herausgenommen, die für mindestens eine der beiden involvierten Variablen keine Informationen liefern. Da Mittelwerte und Varianzen jeweils aus den gültigen Werten der **einzelnen** Variablen berechnet werden können, gibt es den paarweisen Berechnungsmodus in verschiedenen Varianten. Formal hat die »Listwise«-Methode Vorteile, da bei ihr alle Koeffizienten der Korrelationsmatrix auf der gleichen Fallzahl beruhen. Bei paarweisem Ausschluß der Fälle mit fehlenden Daten, kann die Korrelationsmatrix mathematisch unerwünschte Eigenschaften annehmen (nicht mehr »positiv semi-definit« sein), so daß Rechenprogramme (etwa zur multiplen Regression) mit einer Fehlermeldung abbrechen. Allerdings gibt es hierzu wiederum Techniken (wie z. B. das »Glätten« der Korrelationsmatrix), die diesen Defekt aufheben. Schnell (1985), der u. a. die Effizienz von listwise und pairwise deletion unter

verschiedenen experimentellen Bedingungen getestet hat, kommt zu dem Ergebnis, daß sich der paarweise Ausschluß unter fast allen Test-Bedingungen als die leistungsstärkere Methode erwiesen hat (ebd, S. 15). Im Prinzip beruhen aber beide Verfahren auf der MCAR-Annahme.

In zahlreichen Anwendungsfällen ist »Listwise« schon deshalb keine realistische Alternative, weil sie die Fallzahl zu stark reduzieren würde. Bei der Analyse des Abstimmungsverhaltens von Parlamentsabgeordneten zum Beispiel würde die verbliebene Fallzahl regelmäßig gegen Null tendieren, berücksichtigte man nur diejenigen Abgeordneten, die an allen Abstimmungen teilgenommen haben. Außerdem ist stets damit zu rechnen, daß das Fernbleiben bei Abstimmungen nicht zufällig ist. In solchen Situationen müssen fehlende Werte geschätzt werden, um eine Analyse überhaupt durchführen zu können. Die Schätzung kann z. B. mit Hilfe des Regressionsmodells erfolgen: Fehlende Werte einer Abstimmungsvariablen werden mit Hilfe einer Regressionsgleichung geschätzt, in die diejenigen Abstimmungsvariablen (oder ein Teil von ihnen) als Prädiktoren eingehen, die für den betreffenden Fall vollständige Informationen aufweisen. Die Parameter des betreffenden Regressionsmodells werden mit Hilfe der Korrelationsmatrix ermittelt, die man mit einer Variante des paarweisen Ausschlusses zuvor konstruiert hat[4]. Zur Durchführung dieser Analysen eignet sich das Programmsystem BMDP besser als SPSS. Schnell (1988) zeigt allerdings, mit welchen Befehlen die entsprechenden Operationen auch in SPSS durchgeführt werden können.

Eine ausführliche Darstellung und Begründung des Schätzens fehlender Werte mittels Regressionstechniken im Rahmen einer Analyse des Abstimmungsverhaltens von Abgeordneten der Frankfurter Nationalversammlung findet man bei Best/Kuznia (1983). Dieses Beispiel wie auch die analytischen Untersuchungen und Simulationsstudien, über die Schnell (1986) referiert, machen deutlich: eine adäquate Behandlung fehlender Werte ist um so eher zu erwarten, je gründlicher der Mechanismus aufgedeckt wird, der zu fehlenden Informationen geführt hat.

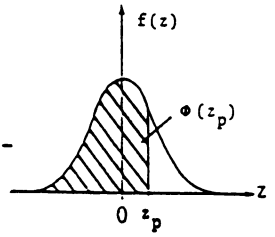
---

[4] Dies kann hier nicht weiter erläutert werden, weil wir in diesem Einführungstext nicht besprochen haben, wie sich die Regressionsanalyse mittels der Matrizenrechnung durchführen läßt.

Anhang A

Tabelle 1 a: Standardnormalverteilung
(Dichtefkt. "von links" integriert)

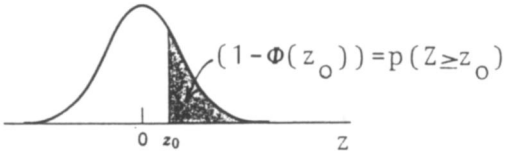
Tabelliert sind für Abszissenwerte z zwischen 0,00 und 3,69 die zugehörigen Werte der Verteilungsfunktion $(\Phi)z$



z_p	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0,1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0,2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0,3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0,4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0,5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0,6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0,7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0,8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0,9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1,0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1,1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1,2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1,3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1,4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1,5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1,6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1,7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1,8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1,9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2,0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2,1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2,2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2,3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2,4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2,5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2,6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2,7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2,8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2,9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3,0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3,1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3,2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995

Quelle: Schaich (1977, S. 319)

Tabelle 1 b: Standardnormalverteilung
(Dichtefkt. "von rechts" integriert)

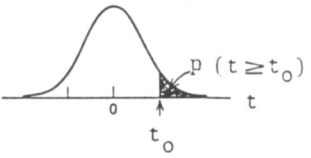


Tabelliert sind für Abszissenwerte z_o zwischen 0,00 und 2,29 die Werte $1-\Phi(z_o)$

z_o	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014

Quelle: Wonnacott/Wonnacott (1972, S. 480)

Tabelle 2: Student's t-Verteilung



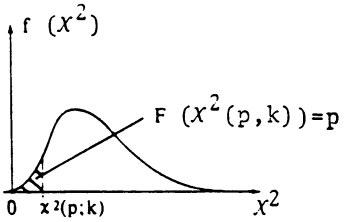
Tabelliert sind Abszissenwerte t_o, die unterschiedlichen Flächensegmenten der integrierten Dichtefunktion bei Vorgabe einer bestimmten Zahl von Freiheitsgraden (d.f.) entsprechen.

df	P = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.253	.674	1 282	1.645	1.960	2.326	2.576	3.090

Quelle: Hays (1973, S. 885)

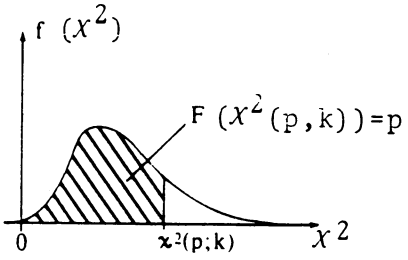
Tabelle 3 a: Chi-Quadrat-Verteilung ("von links" integriert)

Tabelliert sind für ausgewählte Wahrscheinlichkeiten p die zugehörigen Quantile $\chi^2(p,k)$ bei alternativen Anzahlen k von Freiheitsgraden



Anzahl *k* der Freiheitsgrade	Werte *p* der Verteilungsfunktion				
	0,001	0,005	0,025	0,05	0,1
1	—	—	—	0,004	0,016
2	0,002	0,010	0,051	0,103	0,211
3	0,024	0,072	0,216	0,352	0,584
4	0,091	0,207	0,484	0,711	1,06
5	0,210	0,412	0,831	1,15	1,61
6	0,381	0,676	1,24	1,64	2,20
7	0,598	0,989	1,69	2,17	2,83
8	0,857	1,34	2,18	2,73	3,49
9	1,15	1,73	2,70	3,33	4,17
10	1,48	2,16	3,25	3,94	4,87
11	1,83	2,60	3,82	4,57	5,58
12	2,21	3,07	4,40	5,23	6,30
13	2,62	3,56	5,01	5,89	7,04
14	3,04	4,07	5,63	6,57	7,79
15	3,48	4,60	6,26	7,26	8,55
16	3,94	5,14	6,91	7,96	9,31
17	4,42	5,70	7,56	8,67	10,09
18	4,90	6,26	8,23	9,39	10,86
19	5,41	6,84	8,91	10,12	11,65
20	5,92	7,43	9,59	10,85	12,44
21	6,45	8,03	10,28	11,59	13,24
22	6,98	8,64	10,98	12,34	14,04
23	7,53	9,26	11,69	13,09	14,85
24	8,08	9,89	12,40	13,85	15,66
25	8,65	10,52	13,12	14,61	16,47
26	9,22	11,16	13,84	15,38	17,29
27	9,80	11,81	14,58	16,15	18,11
28	10,39	12,46	15,31	16,93	18,94
29	10,99	13,12	16,05	17,71	19,77
30	11,59	13,79	16,79	18,49	20,60
40	17,92	20,71	24,43	26,51	29,05

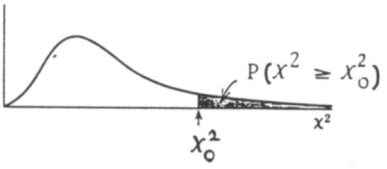
Tabelle 3 a (Forts.)



Anzahl *k* der Freiheitsgrade	Werte *p* der Verteilungsfunktion						
	0,50	0,90	0,95	0,975	0,99	0,995	0,999
1	0,455	2,71	3,84	5,02	6,63	7,88	10,83
2	1,39	4,61	5,99	7,38	9,21	10,60	13,81
3	2,37	6,25	7,81	9,35	11,34	12,84	16,26
4	3,36	7,78	9,49	11,14	13,28	14,86	18,47
5	4,35	9,24	11,07	12,83	15,08	16,75	20,41
6	5,35	10,64	12,59	14,45	16,81	18,55	22,46
7	6,35	12,01	14,06	16,01	18,47	20,28	24,32
8	7,34	13,36	15,51	17,53	20,09	21,96	26,13
9	8,34	14,68	16,92	19,02	21,67	23,59	27,88
10	9,34	15,99	18,31	20,48	23,21	25,19	29,59
11	10,34	17,27	19,67	21,92	24,72	26,76	31,26
12	11,34	18,55	21,03	23,34	26,22	28,30	32,91
13	12,34	19,81	22,36	24,74	27,69	29,82	34,53
14	13,34	21,06	23,68	26,12	29,14	31,32	36,12
15	14,34	22,31	25,00	27,49	30,58	32,80	37,70
16	15,34	23,54	26,30	28,85	32,00	34,27	39,25
17	16,34	24,77	27,59	30,19	33,41	35,72	40,79
18	17,34	25,99	28,87	31,53	34,81	37,16	42,31
19	18,34	27,20	30,14	32,85	36,19	38,58	43,82
20	19,34	28,41	31,41	34,17	37,57	40,00	45,31
21	20,34	29,62	32,67	35,48	38,93	41,40	46,80
22	21,34	30,81	33,92	36,78	40,29	42,80	48,27
23	22,34	32,01	35,17	38,08	41,64	44,18	49,73
24	23,34	33,20	36,42	39,36	42,98	45,56	51,18
25	24,34	34,38	37,65	40,65	44,31	46,93	52,62
26	25,34	35,56	38,89	41,92	45,64	48,29	54,05
27	26,34	36,74	40,11	43,19	46,96	49,64	55,48
28	27,34	37,92	41,34	44,46	48,28	50,99	56,89
29	28,34	39,09	42,56	45,72	49,59	52,34	58,30
30	29,34	40,26	43,77	46,98	50,89	53,67	59,70
40	39,34	51,81	55,76	59,34	63,69	66,77	73,40

Quelle: Schaich (1977, S. 320 f.)

Tabelle 3 b: Chi-Quadrat-Verteilung ("von rechts" integriert)

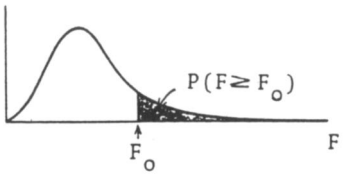


Tabelliert sind die Abszissenwerte x_0^2, die unterschiedlichen Flächensegmenten der integrierten Dichtefunktion bei Vorgabe einer bestimmten Zahl von Freiheitsgraden (d.f.) entsprechen.

P / d.f.	.250	.100	.050	.025	.010	.005	.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100	104	112
80	88.1	96.6	102	107	112	116	125
90	98.6	108	113	118	124	128	137
100	109	118	124	130	136	140	149

Quelle: Wonnacott/Wonnacott (1972, S. 482)

Tabelle 4: F-Verteilung



Tabelliert sind die Abszissenwerte F_o, die unterschiedlichen Flächensegmenten der integrierten Dichtefunktion bei Vorgabe bestimmter Zähler- und Nennerfreiheitsgrade entsprechen.

Nenner-		Freiheitsgrade d. Zählers										
freih.gr.	p	1	2	3	4	5	6	8	10	20	40	∞
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.19	9.32	9.58	9.71	
	.10	39.9	49.5	53.6	55.8	57.2	58.2	59.4	60.2	61.7	62.5	
	.05	161	200	216	225	230	234	239	242	248	251	254
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.35	3.38	3.41	3.45	
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.39	9.44	9.47	
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.5	19.5
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.5	
	.001	998	999	999	999	999	999	999	999	999	999	
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.44	2.44	2.46	2.47	
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.23	5.18	5.16	
	.05	10.1	9.55	9.28	9.12	9.10	8.94	8.85	8.79	8.66	8.59	8.53
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.5	27.2	26.7	26.4	
	.001	167	149	141	137	135	133	131	129	126	125	
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.92	3.84	3.80	
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.80	5.72	5.63
	.01	21.2	18.0	16.7	16.0	15.5	15.2	14.8	14.5	14.0	13.7	
	.001	74.1	61.3	56.2	53.4	51.7	50.5	49.0	48.1	46.1	45.1	
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.88	1.88	
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.30	3.21	3.16	
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.56	4.46	4.36
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.3	10.1	9.55	9.29	
	.001	47.2	37.1	33.2	31.1	29.8	28.8	27.6	26.9	25.4	24.6	
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.77	1.77	1.76	1.75	
	.10	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.94	2.84	2.78	
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	3.87	3.77	3.67
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.10	7.87	7.40	7.14	
	.001	35.5	27.0	23.7	21.9	20.8	20.0	19.0	18.4	17.1	16.4	
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.69	1.67	1.66	
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.70	2.59	2.54	
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.44	3.34	2.71
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.84	6.62	6.16	5.91	
	.001	29.3	21.7	18.8	17.2	16.2	15.5	14.6	14.1	12.9	12.3	
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.63	1.61	1.59	
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.54	2.42	2.36	
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.15	3.04	2.54
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.36	5.12	
	.001	25.4	18.5	15.8	14.4	13.5	12.9	12.0	11.5	10.5	9.92	
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.59	1.56	1.55	
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.42	2.30	2.23	
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	2.94	2.83	3.23
	.01	10.6	8.02	6.99	6.42	6.06	5.80	5.47	5.26	4.81	4.57	
	.001	22.9	16.4	13.9	12.6	11.7	11.1	10.4	9.89	8.90	8.37	
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.56	1.55	1.52	1.51	
	.10	3.28	2.92	2.73	2.61	2.52	2.46	2.38	2.32	2.20	2.13	
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.77	2.66	2.93
	.01	10.0	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.41	4.17	
	.001	21.0	14.9	12.6	11.3	10.5	9.92	9.20	8.75	7.80	7.30	

Tabelle 4 (Forts.)

Nenner freih.gr.	P	Freiheitsgrade d. Zählers 1	2	3	4	5	6	8	10	20	40	∞
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.51	1.50	1.47	1.45	
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.19	2.06	1.99	
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.54	2.43	2.30
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	3.86	3.62	
	.001	18.6	13.0	10.8	9.63	8.89	8.38	7.71	7.29	6.40	5.93	
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.48	1.46	1.43	1.41	
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.10	1.96	1.89	
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.39	2.27	2.13
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.51	3.27	
	.001	17.1	11.8	9.73	8.62	7.92	7.43	6.80	6.40	5.56	5.10	
16	.25	1.42	1.51	1.51	1.50	1.48	1.48	1.46	1.45	1.40	1.37	
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.09	2.03	1.89	1.81	
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.28	2.15	2.01
	.01	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.26	3.02	
	.001	16.1	11.0	9.00	7.94	7.27	6.81	6.19	5.81	4.99	4.54	
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.43	1.42	1.38	1.35	
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.04	1.98	1.84	1.75	
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.19	2.06	1.92
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.08	2.84	
	.001	15.4	10.4	8.49	7.46	6.81	6.35	5.76	5.39	4.59	4.15	
20	.25	1.40	1.49	1.48	1.46	1.45	1.44	1.42	1.40	1.36	1.33	
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.94	1.79	1.71	
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.12	1.99	1.84
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	2.94	2.69	
	.001	14.8	9.95	8.10	7.10	6.46	6.02	5.44	5.08	4.29	3.86	
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.37	1.35	1.30	1.27	
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.82	1.67	1.57	
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	1.93	1.79	1.62
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.55	2.30	
	.001	13.3	8.77	7.05	6.12	5.53	5.12	4.58	4.24	3.49	3.07	
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.35	1.33	1.28	1.24	
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.76	1.61	1.51	
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	1.84	1.69	1.51
	.01	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.37	2.11	
	.001	12.6	8.25	6.60	5.70	5.13	4.73	4.21	3.87	3.15	2.73	
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.32	1.30	1.25	1.21	
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.71	1.54	1.44	
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.75	1.59	1.39
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.20	1.94	
	.001	12.0	7.76	6.17	5.31	4.76	4.37	3.87	3.54	2.81	2.41	
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.30	1.28	1.22	1.18	
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.72	1.65	1.48	1.37	
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.91	1.66	1.50	1.25
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.47	2.03	1.76	
	.001	11.4	7.32	5.79	4.95	4.42	4.04	3.55	3.24	2.53	2.11	
∞												1.00

Quelle: Wonnacott/Wonnacott (1972, S. 484 f.)

Tabelle 5: Fischer's Z-Werte

r	Z	r	Z	r	Z	r	Z	r	Z
0,000	0.000	0,200	0,203	0,400	0,424	0,600	0,693	0,800	1,099
0,005	0,005	0,205	0,208	0,405	0,430	0,605	0,701	0,805	1,113
0,010	0,010	0,210	0,213	0,410	0,436	0,610	0,709	0,810	1,127
0,015	0,015	0,215	0,218	0,415	0,442	0,615	0,717	0,815	1,142
0,020	0,020	0,220	0,224	0,420	0,448	0,620	0,725	0,820	1,157
0,025	0,025	0,225	0,229	0,425	0,454	0,625	0,733	0,825	1,172
0,030	0,030	0,230	0,234	0,430	0,460	0,630	0,741	0,830	1,188
0,035	0,035	0,235	0,239	0,435	0,466	0,635	0,750	0,835	1,204
0,040	0,040	0,240	0,245	0,440	0,472	0,640	0,758	0,840	1,221
0,045	0,045	0,245	0,250	0,445	0,478	0,645	0,767	0,845	1,238
0,050	0,050	0,250	0,255	0,450	0,485	0,650	0,775	0,850	1,256
0,055	0,055	0,255	0,261	0,455	0,491	0,655	0,784	0,855	1,274
0,060	0,060	0,260	0,266	0,460	0,497	0,660	0,793	0,860	1,293
0,065	0,065	0,265	0,271	0,465	0,504	0,665	0,802	0,865	1,313
0,070	0,070	0,270	0,277	0,470	0,510	0,670	0,811	0,870	1,333
0,075	0,075	0,275	0,282	0,475	0,517	0,675	0,820	0,875	1,354
0,080	0,080	0,280	0,288	0,480	0,523	0,680	0,829	0,880	1,376
0,085	0,085	0,285	0,293	0,485	0,530	0,685	0,838	0,885	1,398
0,090	0,090	0,290	0,299	0,490	0,536	0,690	0,848	0,890	1,422
0,095	0,095	0,295	0,304	0,495	0,543	0,695	0,858	0,895	1,447
0,100	0,100	0,300	0,310	0,500	0,549	0,700	0,867	0,900	1,472
0,105	0,105	0,305	0,315	0,505	0,556	0,705	0,877	0,905	1,499
0,110	0,110	0,310	0,321	0,510	0,563	0,710	0,887	0,910	1,528
0,115	0,116	0,315	0,326	0,515	0,570	0,715	0,897	0,915	1,557
0,120	0,121	0,320	0,332	0,520	0,576	0,720	0,908	0,920	1,589
0,125	0,126	0,325	0,337	0,525	0,583	0,725	0,918	0,925	1,623
0,130	0,131	0,330	0,343	0,530	0,590	0,730	0,929	0,930	1,658
0,135	0,136	0,335	0,348	0,535	0,597	0,735	0,940	0,935	1,697
0,140	0,141	0,340	0,354	0,540	0,604	0,740	0,950	0,940	1,738
0,145	0,146	0,345	0,360	0,545	0,611	0,745	0,962	0,945	1,783
0,150	0,151	0,350	0,365	0,550	0,618	0,750	0,973	0,950	1,832
0,155	0,156	0,355	0,371	0,555	0,626	0,755	0,984	0,955	1,886
0,160	0,161	0,360	0,377	0,560	0,633	0,760	0,996	0,960	1,946
0,165	0,167	0,365	0,383	0,565	0,640	0,765	1,008	0,965	2,014
0,170	0,172	0,370	0,388	0,570	0,648	0,770	1,020	0,970	2,092
0,175	0,177	0,375	0,394	0,575	0,655	0,775	1,033	0,975	2,185
0,180	0,182	0,380	0,400	0,580	0,662	0,780	1,045	0,980	2,298
0,185	0,187	0,385	0,406	0,585	0,670	0,785	1,058	0,985	2,443
0,190	0,192	0,390	0,412	0,590	0,678	0,790	1,071	0,990	2,647
0,195	0,198	0,395	0,418	0,595	0,685	0,795	1,085	0,995	2,994

Quelle: Bortz (1979, S. 842)

Anhang B

Das Rechnen mit Erwartungswerten

In Kap. 6 haben wir den Begriff des Erwartungswertes als »Mittelwert« einer Zufallsvariablen eingeführt (siehe Hays 1973, S. 225). Für eine diskrete Zufallsvariable X ist er analog dem arithmetischen Mittel definiert:

$$(A1) \qquad E(X) = \sum_{i=1}^{k} x_i \cdot p(x_i) \quad , \quad \Sigma p(x_i) = 1$$

Dabei stellen die x_i die k Realisationen der Zufallsvariablen X dar und $p(x_i)$ die Wahrscheinlichkeiten, mit denen sie auftreten. Die Definition läßt sich auch allgemeiner fassen, indem sie man sie auf Funktionen g(X) einer Zufallsvariablen X ausdehnt:

$$(A1a) \qquad E[g(X)] = \sum_{i=1}^{k} g(x_i)p(x_i)$$

In Kap. 6.5 haben wir mit $g(X)=[X - E(X)]^2$ eine spezielle Funktion dieser Art eingeführt. Ihr Erwartungswert $E[(X-E(X))^2]$ wird dort als Varianz definiert.

Bei der Definition des Erwartungswertes für stetige (kontinuierliche) Zufallsvariablen X müssen wir auf die Dichtefunktion f(x) zurückgreifen, die über den gesamten Wertebereich $-\infty \leqq x \leqq +\infty$ zu integrieren ist:

$$(A2) \qquad E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \quad , \quad \int_{\infty}^{\infty} f(x)\, dx = 1$$

Allgemein läßt sich analog zu (A1a) definieren:

$$(A2a) \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$$

Erläuterungen hierzu finden sich in Kap. 6.5. In diesem Anhang wollen wir einige Regeln zusammenfassen, nach denen man mit Erwartungswerten Rechenoperationen durchführen kann[1].

(1) Der Erwartungswert einer Konstanten c ist gleich dieser Konstanten:

$E(c) = c$

(2) Wenn c eine reelle Zahl und X eine Zufallsvariable mit dem Erwartungswert E(X) ist, gilt

(2a) $E(cX) = c \cdot E(X)$
(2b) $E(X + c) = E(X) + c$

(3) Wenn X und Y Zufallsvariablen mit dem Erwartungswert E(X) resp. E(Y) sind, dann ist

$$E(X + Y) = E(X) + E(Y)$$

Diese Regel kann auf eine beliebige endliche Reihe von Zufallsvariablen $X_1, X_2, \ldots, X_K$ ausgedehnt werden:

$$E(X_1 + X_2 + \ldots + X_K) = E(X_1) + E(X_2) + \ldots + E(X_k)$$

Die Regel ist auch dann anwendbar, wenn eine dieser Variablen die Funktion einer anderen Variablen ist, z. B.

$$X_3 = 4 \cdot X_1^2$$

(4) Falls eine endliche Menge von Zufallsvariablen $\{X_K\}$ mit $k = 1,2, \ldots K$ gegeben ist, die alle voneinander unabhängig sind, so gilt

$$E(X_1 \cdot X_2 \cdot \ldots \cdot X_K) = E(X_1) \cdot E(X_2) \cdot \ldots \cdot E(X_K)$$

Der Erwartungswert des Produkts der unabhängigen Zufallsvariablen ist gleich dem Produkt der Erwartungswerte.

Als Beispiel für die Anwendung dieser Rechenregeln wollen wir das in Kap. 6, Gleichung (6-37) eingeführte Theorem beweisen, wonach die Varianz einer Zufallsvariable X mit $E(X^2) - [E(X)]^2$ gegeben ist.
Laut Definition (6-36) ist

---

[1] Eine ausführliche Darstellung einschließlich Beweisen findet man beispielsweise in Hays (1973, S. 871ff.) oder Bortz (1979, S. 790ff.).

(A3)

$V(X) = E[X-\mu]^2 \quad , \mu = E(X)$

Durch Ausmultiplizieren des Binoms erhält man

(A3a)

$V(X) = E[X^2 - 2X\mu + \mu^2]$

Nach Regel 3 ergibt sich daraus

(A3b)

$V(X) = E(X^2) - E(2X\mu) + E(\mu^2)$

Da μ und μ^2 ebenso wie der Faktor 2 eine Konstante sind, wird daraus nach Regel 2a

(A3c)

$V(X) = E(X^2) - 2\mu E(X) + \mu^2$

Setzen wir wieder $\mu = E(X)$ ein, so erhalten wir

(A3d)

$V(X) = E(X^2) - 2[E(X)E(X)] + [E(X)]^2$

An dieses Ergebnis anknüpfend, können wir auch das Theorem (6-38) beweisen, wonach $V(a+bX) = b^2V(X)$, wenn X eine Zufallsvariable ist und a und b Konstanten aus dem reellen Zahlenbereich sind.

Laut Definition der Varianz und dem Ergebnis in Gleichung (A3d) ist

(A4)

$$V(a+bX) = E[(a+bX) - E(a+bX)]^2 \text{ , siehe (6-36)}$$
$$= E[(a+bX)^2] - [E(a+bX)]^2 \text{ , siehe (A3a)}$$

(Dabei machen wir von dem Satz Gebrauch, daß die Lineartransformation einer Zufallsvariablen, z.B. Y = a+bX, wiederum eine Zufallsvariable ergibt.)

$$= E[a^2 + 2abX + (bX)^2] - [a + bE(X)]^2$$

$$= a^2 + 2abE(X) + b^2E(X^2)$$

$$\quad - a^2 - 2abE(X) - b^2[E(X)]^2$$

$$= b^2E(X^2) - b^2[E(X)]^2$$

$$= b^2[E(X^2) - (E(X))^2]$$

$$= b^2\, V(X)$$

LITERATURVERZEICHNIS

Achen, Christopher H., Interpreting and using regression, Beverly Hills, London, New Delhi 1982 (Series: Quantitative Applications in the Social Sciences 29).

Aldrich, John H./**Nelson**; Forrest D., Linear probability, logit and probit models, Beverly Hills, London, New Delhi 1984 (Series: Quantitative Applications in the Social Sciences 45).

Allison, D., Testing for interaction in multiple regression, in: American Journal of Sociology 83 (1977), S. 144-153.

Althauser, Robert P., Multicollinearity and non-additive regression models, in:

Blalock, Hubert M. (Hg.), Causal models in the social sciences, Chicago 1971, S. 453-472.

Anderson, Andy B./**Basilevsky**, Alexander/**Hum**, Derek P.J., Missing data, in: Peter H. **Rossi**/James D. **Wright**/Andy B. **Anderson** (Hrsg.), Handbook of survey research, New York usw. 1983, S. 415-494.

Asher, Herbert B., Causal modelling, Beverly Hills, London, New Delhi 1983, 2nd edition (Series: Quantitative Applications in the Social Sciences 3).

Belsley, David A./**Kuh**, Edwin/**Welsch**, Roy E., Regression diagnostics: identifying influential data and sources of collinearity, New York usw. 1980.

Berk, Richard A., An introduction to sample selection bias in sociological data, in: American Sociological Review 48 (1983), S. 386-398.

Berry, William D./**Feldmann**, Stanley, Multiple regression in practice, Beverly Hills, London, New Delhi 1985 (Series: Quantitative Applications in the Social Sciences 50).

Best, Heinrich/**Kuznia**, Reiner, Die Behandlung fehlender Werte bei der seriellen Analyse namentlicher Abstimmungen, oder: Wege zur Therapie des Horror Vacui, in: Historical Social Research 26 (1983), S. 49-82.

Blalock, Hubert M., Social Statistics, New York usw. 1960.

Blalock, Hubert M., Causal inferences, closed populations, amd measures of association, in: Ders. (Hrsg), Causal models in the social sciences, Chicago 1971, S. 139-151.

Böltgen, Ferdinand, Auswahlverfahren. Eine Einführung für Sozialwissenschaftler, Stuttgart 1976.

Bortz, Jürgen, Lehrbuch der Statistik. Für Sozialwissenschaftler, Berlin, Heidelberg, New York 1979.

Bollen, Kenneth A., Structural equations with latent variables, New York usw. 1989.

Cook, R.D./**Weisberg**, S., Residuals and influence in regression, New York 1982.

Darlington, Richard B., Multiple regression in psychological research and practice, in: Psychological Bulletin 59 (1968), S. 161-182.

Draper, N.R./**Smith**, H., Applied regression analysis, New York usw. 1981 (2nd edition).

Floud, Roderick, Einführung in quantitative Methoden für Historiker, Stuttgart 1980.

Friedrich, Robert J., In defense of multiplicative terms in multiple regression equations, in: American Journal of Political Science 26 (1982), S. 797-833.

FU-Autorenkollektiv, Skript zur Statistischen Grundausbildung Teil I + II, 1976.

Hanushek, Eric A./**Jackson**, John E., Statistical methods for social scientists, Orlando usw. 1977.

Hartung, Joachim/**Elpelt**, Bärbel/**Klösener**, Karl-Heinz, Statistik. Lehr- und Handbuch der angewandten Statistik, München, Wien 1986 (5. Auflage).

Hays, William L., Statistics for the social sciences, London usw. 1973 (2nd edition).

Hensher, David A./**Johnson**, Lester W., Applied discrete choice modelling, New York 1981.

Hoaglin, D.C./**Mosteller**, F./**Tukey**, J.W., Exploring data tables, trends and shapes, New York 1985.

Holm, Kurt, Die Befragung 5. Pfadanalyse, Coleman-Verfahren, München 1977.

Hummel, H.J., Probleme der Mehrebenenanalyse, Stuttgart 1972.

Jagodzinski, Wolfgang/**Weede**, Erich, Testing curvi-linear propositions by polynomial regression with particular reference to the interpretation of standardized solutions, in: Quality and Quantity 15 (1981), S. 447-463.

Jarausch, K.H./**Arminger**, G./**Thaller**, M., Quantitative Methoden in der Geschichtswissenschaft. Eine Einführung in die Forschung, Datenverarbeitung und Statistik, Darmstadt 1985.

Kendall, Maurice K./**Stuart**, Alan/**Keith**, J., Kendall's advanced theory of statistics, Vol. 1: Distribution theory, London 1987 (5th edition).

Kerlinger, Fred N./**Pedhazur**, E.J., Multiple regression in behavioral research, New York usw. 1973.

Kim, Jae-On/**Curry**, James, The treatment of missing data in multivariate analysis, in: Sociological Methods & Research 6 (1977), S. 215-240.

Kim, Jae-On/**Ferree**, G.D., Standardization in causal analysis, in: Sociological Methods & Research 10 (1981), S. 187-210.

King, Gary, How not to lie with statistics: avoiding common mistakes in quantitative political science, in: American Journal of Political Science 30 (1980), S. 666ff.

Kirschner, Hans-Peter, ALLBUS 1980: Stichprobenplan und Gewichtung, in: K.U. **Mayer**/ P. **Schmidt** (Hrsg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, Frankfurt a.M., New York 1984, S. 114-182.

Kmenta, Jan, Elements of econometrics, New York, London 1971.

Kregel, Ulrich, Einführung in die Wahrscheinlichkeitstheorie und Statistik, Braunschweig 1988.

Kriz, Jürgen, Statistik in den Sozialwissenschaften, Reinbek bei Hamburg 1973.

Kühnel, Steffen, Kausale Effekte oder Varianzzerlegung? Eine Anmerkung zu Dieter Holtmanns »Interpretation der Effekte in der multivariaten Modellbildung«, in: Zeitschrift für Soziologie 14 (1985), S. 247-248.

Kühnel, Steffen/**Jagodzinski**, Wolfgang/**Terwey**, Michael, Teilnehmen oder Boykottieren: Ein Anwendungsbeispiel der binären logistischen Regression mit SPSSx, in: ZA-Information 25 (1989), S. 44-75.

Lamm, Doron, British soldiers of the first World War: creation of a representative sample, in: Historical Social Research 13/4 (1988), S. 55-98.

Liebetrau, Albert M., Measures of association, Beverly Hills, London, New Delhi 1983 (Series: Quantitative Applications in the Social Sciences 32).

Linder, Arthur/**Berchtold**, Willi, Statistische Auswertung von Prozentzahlen. Probit- und Logitanalyse mit EDV, Basel, Stuttgart 1976.

Marsden, Peter V., Conditional effects in regression models, in: Ders. (Hrsg.), Linear models in social research, Beverly Hills, London 1981, S. 97-116.

Miller, Michael K./**Farmer**, Frank L., Substantive nonadditivity in social science research. A note on induced collinearity and measurement and testing on effects, in: Quality and Quantity 2 (1988), S. 221-237.

Mood, Alexander M./**Graybill**, Franklin A./**Boes**, Duane C., Introduction to the the theory of statistics, McGraw-Hill 1974 (3rd edition).

Norusis, Marija J., SPSSx Advanced Statistics Guide, New York usw. 1985.

Opp, Karl-Dieter/**Schmidt**, Peter, Einführung in die Mehrvariablenanalsyse. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen. Reinbek bei Hamburg 1976.

Pfeifer, Andreas/**Schmidt**, Peter, LISREL. Die Analyse komplexer Strukturgleichungsmodelle, Stuttgart, New York 1987.

Pindyck, Robert S./**Rubinfeld**, Daniel L., Econometric models and econometric forecasts, New York usw. 1981 (2nd edition).

Rochel, Hubertus, Planung und Auswertung von Untersuchungen im Rahmen des allgemeinen linearen Modells, Berlin usw. 1983.

Rohlinger, Harald, Quellen als Auswahl - Auswahl aus Quellen, in: Historical Social Research 24 (1982), S. 34-62.

Schaich, Eberhard, Schätz- und Testmethoden für Sozialwissenschaftler, München 1977. Schlittgen, Rainer, Einführung in die Statistik. Analyse und Modellierung von Daten, München, Wien 1987 (Neuauflage 1990).

Schnell, Rainer, Realization of missing data techniques within statistical programm packages and their empirical performance. Paper to be presented on the Cologne Computer Conference on September 8th, 1988.

Schnell, Rainer, Zur Effizienz einiger Missing-Data-Techniken - Ergebnisse einer Computer-Simulation, in: ZUMA-Nachrichten 17 (1985), S. 50-74.

Schnell, Rainer, Missing-Data-Probleme in der empirischen Sozialforschung, Diss. Bochum 1986.

Schnell, Rainer/**Hill**, Paul B./**Esser**, Elke, Methoden der empirischen Sozialforschung, München, Wien 1988 (Neuauflage 1989).

Shapiro, Gilbert/**Markoff**, John/**Duncan Baretta**, Silvio R., The selective transmission of historical documents: the case of the Parish Cahiers of 1789, in: Histoire & Mesure II,3/4 (1987), S. 115-172.

Siegel, Sidney, Nonparametric statistics for the behavioral sciences, New York usw. 1956.

Southwood, Kenneth E., Substantive theory and statistical interaction: Five models, in: American Journal of Sociology 83 (1978), S. 1154-1203.

SPSS INC., SPSSx User's Guide. 3rd edition. Chicago 1988.

SPSS INC., SPSSx Advanced Statistics Guide. Chicago 1985.

Stolzenberg, Ross M., The measurement and decomposition of causal effects in nonlinear and nonadditive models, in: Karl F. **Schuessler** (Hg.), Sociological Methodology 1980, San Francisco usw. 1979, S. 459-48.

Stoto, Michael A./**Emerson**, John D., Power transformations for data analysis, in: Samuel **Leinhardt** (Hg.), Sociological Methodology 1983-1984, San Francisco usw. 1983, S. 126-168.

Tufte, Edward R., Data analysis for politics and policy, Englewood Cliffs, N.J. 1974.

Wonnacott, Thomas H./**Wonnacott**, Ronald J., Introductory statistics, New York usw. 1972 (2nd edition). von Wright, Georg Henrik, Erklären und Verstehen, Frankfurt a. M. 1974.

Register

Korrekturen

S. 11: In der zweiten Spalte der Elementarereignisse ist die zweite bis vierte Zeile irrtümlich aus der ersten Spalte kopiert worden. Die zweite Spalte muß folgende Elemente enthalten: (mmw) (mwm) (wmm) (mmm)

S. 17: Definitionsgleichung (6-6): Das erste A ist mit einem Querstrich zu versehen: $\bar{A} = \Omega\text{-}A$.

S. 18: Gleichung (6-13): Der letzte Ausdruck ist nicht 33/88, sondern 38/88.
Am Ende des mittleren Absatzes fehlt der Hinweis: "hier als Abb. 6.3".

S. 19: 4. Zeile von oben: Statt "disjunkte Ereignisse" ist "nicht-disjunkte Ereignisse" zu schreiben.

S. 26: Textabschnitt unter Gleichung (6-28): Der in der Mitte stehende Satz wird wie folgt umformuliert: "Er läßt sich anschaulich wie folgt interpretieren: Würde man vor der Durchführung des Zufallsexperiments diesen (Erwartungs-)Wert als Ergebnis des Zufallsexperiments voraussagen, so würde diese Prognose auf lange Sicht (bei "sehr vielen Versuchen") zur geringsten Fehlerquadratsumme führen."

S. 38: 2. Absatz, 4. Zeile: Die Ziffer "5" ist durch "4" zu ersetzen.
2. Absatz, 7. Zeile: Die Ziffer "1,6" ist durch "6,2" zu ersetzen.

S. 39: 7. Zeile von oben: Die zweite Wahrscheinlichkeitsangabe muß lauten: $P(\bar{A}) = (1\text{-}p)$.

S. 41: Die Fußnotenziffer "2" am Ende des 1. Absatzes ist an anderer Stelle zu plazieren, damit sie nicht irrtümlich als Exponent gelesen werden kann.

S. 44: Gleichung (7-6): Im Exponenten fehlt das negative Vorzeichen.

S. 47: Im Exponenten der Dichtefunktion fehlt das negative Vorzeichen.

S. 63: Abb. 7.6 b: Das arithmetische Mittel ist nicht 950, sondern 900.

S. 70: Die rechte Seite der Gleichung (8-3) lautet: $6{,}30 \pm 1{,}241$.
7. Zeile von unten: Es muß heißen: "das 97,5-%-Quantil".

S. 74: Gleichung (8-11), 2. Gleichungszeile: Das Wurzelzeichen gilt für die Summe $(0{,}95 + 0{,}744)$, nicht nur für den ersten Summanden.

S. 78: Gleichung (8-17): Es fehlt eine rechte äußere Klammer nach dem Exponenten "2"; der Erwartungsoperator bezieht sich auf den quadrierten Ausdruck.

S. 79: 4. Zeile über Gleichung (8-21): Das Summenzeichen Σ ist durch π zu ersetzen.

S. 81: Gleichung (8-24): Das "x" im Index von "z" ist mit einem Querstrich zu versehen, um das arithmetische Mittel zu kennzeichnen.

S. 101: Abschnitt 8.7.3 Ende des ersten Absatzes: Das Symbol "ß" ist durch "ρ" zu ersetzen.

S. 125: Gleichung (10-11): Der erste Ausdruck muß lauten: $b_{yx} = r(s_y/s_x)$.

S. 139: Absatz unter Ziff. 6: Dem Symbol "y_i" ist das Attribut "bedingt" (bedingter ...) hinzuzufügen, da es um die Verteilung der abh. Variable um die Regressionsgerade geht.

S. 151: Der Wahlkreis heißt nicht "Empen Achen", sondern "Eupen-Aachen".

S. 166: Vierte Zeile unterhalb der Gleichung (11-8): Der Standardfehler ist "$\sigma(b_1)$".

S. 192: Dritte Zeile unter COMPUTE GEWLOG: Statt "kleiner 0" muß es heißen "kleiner/gleich 0"

S. 220: Fußnote, zweitletzte Zeile: Es muß heißen: "Basis a", nicht "Basis a_e".

S. 232: Fußnote 1. Zeile: Es muß heißen: $P = \binom{n}{a}\pi^a(1-\pi)^{n-a}$.

S. 271: Im Literaturverzeichnis haben die Arbeiten "Schlittgen, Rainer" und "von Wright, Georg Henrik" keine eigenen Absätze erhalten, sondern sind irrtümlich in die Angaben zu "Schaich" bzw. "Wonnacott" einbezogen worden.